Abstract representations of attributed emotion: evidence from neuroscience and development

A DISSERTATION PRESENTED
BY
AMY ELIZABETH SKERRY
TO
THE DEPARTMENT OF PSYCHOLOGY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF PSYCHOLOGY

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
MAY, 2015
Abstract representations of attributed emotion: evidence from neuroscience and development

Abstract

Humans can recognize others’ emotions based on overt cues such as facial expressions, affective vocalizations, or body posture, or by recruiting an abstract, causal theory of the conditions that tend to elicit different emotions. Whereas previous research has investigated the recognition of emotion in specific perceptual modalities (e.g. facial expressions), this dissertation focuses on the abstract representations that relate observable reactions to their antecedent causes. A combination of neuroimaging, behavioral, and developmental methods are used to shed light on the mechanisms that support various forms of emotion attribution, and to elucidate the core features or dimensions that structure the space of emotions we represent.

Chapter 1 identifies brain regions that contain information about emotional valence conveyed either via facial expressions and or via animations depicting abstract situational information. These data reveal regions with modality-specific representations of emotional valence (i.e. patterns of activity that discriminate only positive versus negative facial expressions), as well as modality-independent representations: in medial prefrontal cortex (MPFC), the valence representation generalizes across stimuli, indicating a common neural code that abstracts away from specific perceptual features and is invariant to different forms of evidence. Building on evidence that young infants discriminate and respond to the emotional expressions of others, Chapter 2 investigates whether infants also represent these expressions in relation to the situations that elicit them. The results of several experiments demonstrate that infants within their first year of life have expectations about how facial and vocal
displays of emotion relate to the valence of events that precede them. Whereas Chapters 1 and 2 focus on a simple binary distinction between positive and negative affect, Chapter 3 investigates a space of more fine-grained discriminations (e.g. someone feeling proud vs. grateful). A combination of multi-voxel pattern analyses and representational similarity analyses reveal brain regions containing abstract and high-dimensional representations of attributed emotion. Moreover, a set of causal features (encoding properties of eliciting events that vary between different emotions) outperforms more primitive dimensions in capturing neural similarities within these regions. Together, these studies provide a newly detailed characterization of the representations that structure emotion attribution, including their development and neural basis.
Table of Contents

0. General Introduction

0.1 INTRODUCTION ................................................................. 1
0.2 PERCEIVING EMOTIONS ......................................................... 3
  0.2.1 Adult perception of facial expressions .............................. 4
  0.2.2 The development of facial emotion recognition .................. 5
  0.2.3 Neural mechanisms for emotion perception ...................... 7
0.3 THE PRESENT RESEARCH: INFERRING EMOTIONS ..................... 10
  0.3.1 Chapter 1 ................................................................. 12
  0.3.2 Chapter 2 ................................................................. 14
  0.3.3 Chapter 3 ................................................................. 16
0.4 CONCLUSIONS ................................................................. 17

1. A common neural code for perceived and inferred emotion
   in collaboration with Dr. Rebecca Saxe

1.0 ABSTRACT ........................................................................... 18
1.1 INTRODUCTION ................................................................. 19
1.2 MATERIALS AND METHODS ............................................... 20
1.3 RESULTS ........................................................................... 32
  1.3.1 Experiment 1 .............................................................. 32
  1.3.2 Experiment 2 .............................................................. 39
1.4 DISCUSSION ...................................................................... 40

2. Preverbal infants identify emotional reactions that are incongruent with goal
   outcomes
   in collaboration with Dr. Elizabeth Spelke

2.0 ABSTRACT ........................................................................... 45
2.1 INTRODUCTION ................................................................. 46
2.2. EXPERIMENT 1 ................................................................. 48
  2.2.1 Method ................................................................. 49
  2.2.2 Results ................................................................. 53
  2.2.3 Discussion .............................................................. 54
2.3. EXPERIMENT 2 ................................................................. 55
  2.3.1 Method ................................................................. 56
  2.3.2 Results ................................................................. 57
  2.3.3 Discussion .............................................................. 58
2.4. EXPERIMENT 3 ................................................................. 59
  2.4.1 Method ................................................................. 60
  2.4.2 Results ................................................................. 62
  2.4.3 Discussion .............................................................. 63
2.5. GENERAL DISCUSSION
3. Neural representations of emotion are organized around abstract event features
   in collaboration with Dr. Rebecca Saxe

   3.0 ABSTRACT .......................................................... 71
   3.1 INTRODUCTION ..................................................... 72
   3.2 RESULTS .......................................................... 78
   3.3 DISCUSSION ....................................................... 87
   3.4 EXPERIMENTAL PROCEDURES ................................. 92
   3.5 SUPPLEMENTAL INFORMATION .................................. 99
       3.5.1 Supplemental Experimental Procedures .................. 99

4. General Conclusion

   4.1 INTRODUCTION .................................................. 106
   4.2 SUMMARY AND DISCUSSION .................................... 106
       4.2.1 Chapter 1 Summary ........................................... 106
       4.2.2 Limitations & Open Questions ............................ 107
       4.2.3 Chapter 2 Summary ........................................... 109
       4.2.4 Limitations and Open Questions .......................... 111
       4.2.5 Chapter 3 Summary ........................................... 113
       4.2.6 Limitations and Open Questions .......................... 114
   4.3 MOTIVATING THE APPROACH ................................... 116
       4.3.1 Integrating neuroscience and developmental science .... 117
       4.3.2 Emotion inference as a case study ........................ 119
   4.4 CHALLENGES FOR FUTURE RESEARCH .......................... 128
   4.5 FINAL CONCLUSIONS ............................................ 130

Appendix I. Chapter 1 Supplementary Results ............................ 131
Appendix II. Chapter 3 Supplementary Results and Figures .................... 132
References ................................................................. 141
Acknowledgements

This thesis would not have been possible without the ideas, advice, support, and patience of many people, but most importantly my advisors and collaborators Elizabeth Spelke, Susan Carey, and Rebecca Saxe.

I want to thank Susan for her brilliant and deeply selfless advising. I have never encountered in another scientist such an unadulterated passion for the questions and theoretical issues. While not a formal collaborator on any of the work in this dissertation, Susan’s ideas run through many corners of it.

I want to thank Liz for her boundless enthusiasm, and her ability to find what is exciting in any bit of data. Liz has an amazing creativity for experiments— I’m not sure she has ever considered a question without immediately designing an elegant set of conditions for testing it. I learned an enormous amount in trying to tackle hard problems with Liz.

I want to thank Rebecca for constantly challenging and inspiring me intellectually. I find myself thinking clearer, deeper thoughts after every conversation with her. Rebecca showed me it was possible to be critical, rigorous, and theoretically ambitious, while staying optimistic and enthused about the challenges of this field. I am also deeply grateful to Rebecca for her mentorship and kindness. I would not be where I am today without the personal and intellectual support I received in her lab.

I want to thank all three of my advisors for their patience and respect throughout the last five years. The freedom to explore and learn and periodically pause to consider next steps has been immensely valuable.

I want to thank my final committee member, Nancy Kanwisher, for being a scientific inspiration on every level.

Many additional thanks go to the various friends from Saxe lab, Spelke lab, and Carey lab who have supported me, challenged me, and taught me along the way. Particular appreciation goes to Todd Thompson, Jorie Koster-Hale, and Michael Cohen for essential science and non-science conversations throughout the last 5 years, and to Lindsey Powell for being a role model and mentor from my first earliest days in Spelke Lab. Finally, thanks to RJ Ryan for boundless patience and support. You are selfless and wonderful.
“SHE WAS A GENIUS OF SADNESS, immersing herself in it, separating its numerous strands, appreciating its subtle nuances … Brod discovered 613 sadesses, each perfectly unique, each a singular emotion, no more similar to any other sadness than to anger, ecstasy, guilt, or frustration. Mirror Sadness. Sadness of Domesticated Birds. Sadness of Being Sad in front of One’s Parent. Humor Sadness. Sadness of Love Without Release.”

-Jonathan Safran Foer, Everything Is Illuminated
General Introduction

0.1 INTRODUCTION

The ability to recognize the emotional states of other people is central to many of our species unique social and cooperative behaviors. Attributing emotions to others allows us to explain their actions (Graham, 1988; Harris, 1989), to predict their future behavior (Repacholi & Gopnik, 1997; Vander Wyk, Hudac, Carter, Sobel, & Pelphrey, 2009), and to learn about the world efficiently by capitalizing on the experiences of others (Campos, 1983; Mumme, Fernald, & Herrera, 1996; Striano & Rochat, 2000). Effectively reasoning about others’ emotions is related to successful social functioning across the lifespan (Denham et al., 2003; Eisenberg & Fabes, 1990; Frederickson, Petrides, & Simmonds, 2012; Scherer & Scherer, 2011; Vaish, Carpenter, & Tomasello, 2009), and impaired emotion attribution is a core deficit in a number of psychopathologies (e.g. Autism Spectrum Disorders: Adolphs, Sears, & Piven, 2001; Baron-Cohen, et al., 1997; Harms, Martin, & Wallace, 2010; Hobson, 1986; Schizophrenia: Amminger et al., 2012; Gold et al., 2012; psychopathy: de Wied, Gispen-de Wied, & van Boxtel, 2010; Fairchild et al., 2009;2010). The ability to recognize and understand emotions has also attracted attention in the fields of artificial intelligence and machine learning, as researchers strive to create emotionally intelligent interfaces across a range of applications (Bediou et al., 2012; Dadds, Cauchi, Wimalaweera, Hawes, & Brennan, 2012; Stanković, Nešić, Obrenović, Stojanović, & Milošević, 2015).

How do we recognize the emotional states of others? In this thesis, I investigate the neural and cognitive mechanisms that support emotion attribution, as well as their origins in human development. Importantly, there
are diverse sources of evidence available for identifying someone’s emotions: we can recognize what others are feeling based on their facial expressions (Ekman, 1992; Izard, 1971; Posamentier & Abdi, 2003), affective vocalizations or prosody (Bachorowski & Owren, 2003; Sauter, Eisner, Ekman, & Scott, 2010; Scherer, 2003), and gait or body posture (Atkinson, Dittrich, Gemmell, & Young, 2004; Aviezer, Trope, & Todorov, 2012; Dael, Mortillaro, & Scherer, 2012; de Gelder, 2006). However, we can also attribute emotions based solely on the context or situation a person is in by recruiting abstract conceptual knowledge about emotions (de Vega, Díaz, & León, 1997; Ortony, 1990; Parkinson, 2007; Siemer & Reisenzein, 2007). In this thesis, I argue that such knowledge takes the form of an intuitive theory that captures regularities in the conditions that elicit different emotions; this theory allows us to reason about a range of diverse and subtle emotional states that extend beyond the set of emotions associated with canonical emotional displays (see Fontaine, Scherer, Roesch, & Ellsworth, 2007a; Frijda, 1986; Ortony & Turner, 1990).

Compared to perceptual emotion recognition, the ability to infer emotional states from antecedent situations or events has received minimal attention in cognitive science and neuroscience. Relatively little is known about how conceptual knowledge of others’ emotions is acquired and organized, or how that knowledge is encoded the human brain. The following three chapters aim to fill this gap.

Emotion attribution also provides a rich case study for investigating fundamental questions in cognitive science. First, because emotions can be identified from a range of qualitatively different sources, emotion attribution serves as an example of high-level abstraction and invariance (see DiCarlo & Cox, 2007; DiCarlo, Zoccolan, & Rust, 2012; Ullman, 1998). How does the brain construct a representation of a concept like “sadness” that is invariant to the diverse inputs that can give rise to it (e.g. an upset facial expression, a quivering voice, a description of a distressing event)? An important step towards tackling this question would be to characterize the content and level of abstraction of emotion representations in the different brain regions involved in transforming perceptual inputs into high-level emotion concepts.

Second, understanding others’ emotions often requires that a number of distinct representations (e.g. of a goal/desire, a goal outcome, and an emotional display) be integrated into a single coherent theory. What
mechanisms support such theory-like inferences, and how do the necessary representational resources emerge over human development (Carey, 1985, 2009; Gopnik & Wellman, 1994)? The domain of emotion provides a novel case study for understanding how rich, flexible, causal theories could be constructed using the representational primitives of potentially more limited innate systems (Spelke and Kinzler, 2007).

Finally, a defining goal of cognitive science is to characterize the content and organization of representations involved in specific domains of perception and cognition (e.g. the representations underlying object recognition: Biederman, 1987; Buehlof, Edelman, & Tarr, 1994; Riesenerhuber & Poggio, 1999; Tarr & Bülthoff, 1998; Ullman, 2007). Recent advances in visual neuroscience show that it is possible to characterize fine-grained features of representations in different brain regions and at different stages of processing (e.g. the parameters that structure the hierarchy of object representations in inferotemporal (IT) cortex: Cox & DiCarlo, 2008; DiCarlo et al., 2012; Lehky, et al., 2014; Sato et al., 2013; Yamins et al., 2014). While this approach has been successful for visual perception, the representations underlying abstract, high-level reasoning processes (e.g. causal inference, theory of mind) have remained elusive. Can we not only identify the neural mechanisms recruited when attributing emotions to others, but also to characterize the underlying representations that support these theory-like inferences? A primary aim of this thesis is to show that neural and behavioral methods can be used to jointly characterize the fine-grained representational structure of a high-level domain like emotion attribution.

In the following chapters, I report three lines of research on the mechanisms by which we infer the emotional states of others, shedding light on both their neural basis and their developmental origins. In this introductory chapter, I begin with a review of the prior literature on emotion perception, focusing particularly the attribution of emotion from overt facial displays. I then discuss the aims of each of the following three chapters. In each chapter, I aim to 1) show that humans make abstract attributions of emotion in the absence of any perceptual cues, and 2) bring data to bear on the representations that support these situation-based inferences.

0.2 PERCEIVING EMOTIONS
Throughout this thesis, I argue that the study of facial expressions provides a fundamentally limited window into emotion knowledge. Still, research on facial emotion recognition is rich and extensive (Adolphs, 2002b; Blair, 2003; Ekman, 1993; Posamentier & Abdi, 2003; Shariff & Tracy, 2011; Vuilleumier & Pourtois, 2007), and frames the theoretical landscape relevant to the current work. Moreover, this research shows that it is possible to characterize specific perceptual invariants that allow for the identification of emotion in different perceptual modalities, and has yielded insights into how these perceptual representations develop, and how they are implemented in the human brain. The overarching goal of the research reported in following three chapters is to provide a similarly rich characterization of the abstract event regularities that allow for identification of emotions from descriptions of eliciting situations, and to sketch out a preliminary picture of how this intuitive knowledge is encoded in neural populations, and how it emerges in human development.

0.2.1 Adult perception of facial expressions

The perception of emotion from overt expressions has been studied for decades (Abelson & Sermat, 1962; Engen, Levy, & Schlosberg, 1958; Jarden & Fernberger, 1926; Landis, 1924; Thompson & Meltzer, 1964), and conflicting theories on the structure, dimensions, and universality of facial emotion displays have formed one of the longest-standing debates in modern psychology (Barrett, 2011; Ekman & Cordaro, 2011; Levenson, 2011; Lindquist, Siegel, Quigley, & Barrett, 2013). According to basic emotion theory, there are a set of 5 or 6 discrete emotional states, each a distinct evolutionary adaptation with an innate physiological response and universally recognizable display (Ekman, 1992; Ekman & Rosenberg, 1997; Izard, 1971; Panksepp, 1992). These basic, universal emotions are argued to provide the building blocks for constructing more complex emotional expressions (Du, Tao, & Martinez, 2014; Ekman & Cordaro, 2011). An alternative proposal, termed the “circumplex model”, posits instead only two core dimensions—valence and arousal—that are reliably extracted from facial expressions (Barrett, 2006b; Russell, 1980; Russell & Bullock, 1986a, 1986b; Watson, Wiese, Vaidya, & Tellegen, 1999). On this view, the space of universal primitives is not only smaller, but also based on continuous dimensions rather than categories (Gendron, Roberson, van der Vyver, & Barrett, 2014; Jack, Caldara, & Schyns, 2012; Jack, Garrod, Yu, Caldara, & Schyns, 2012; Posner et al., 2009; Russell, 1994).
Outside this debate, recent research has begun to characterize how emotion-relevant facial information is detected and perceived (Adolphs, 2002a, 2002b; Martinez and Du, 2012; Posamentier & Abdi, 2003; Vuilleumier et al., 2003). Emotional faces can be processed rapidly and even unconsciously (Dimberg, 1997; Pessoa, 2005; Pourtois, Grandjean, Sander, & Vuilleumier, 2004; Smith, 2011): what visual regularities allow for efficient extraction of emotion from faces? A number of psychophysical studies suggest that emotions are recognized using both configural and featural information (Bimler & Paramei, 2006; Bimler et al., 2013; Bombari et al., 2013; Calder et al., 2000; Oosterhof & Todorov, 2009; Said, Sebe, & Todorov, 2009). Facial motion seems to be a particularly dominant informational source (e.g. recognition advantage for dynamic over static expressions: Ambadar, Schooler, and Cohn, 2005; Bassili, 1978, 1979), especially motion of the eye and mouth regions (Ambadar et al., 2005; Adolphs et al., 2005; Ekman & Rosenberg, 1997).

A particularly fruitful approach has been to parameterize different facial expressions in terms of a large set of anatomy-based Facial Action Units (or the Facial Action Coding System: FACS: Ekman & Rosenberg, 1997). Decomposing complex expressions into additive and non-additive combinations of muscular action primitives has revealed a number of specific perceptual invariants that support emotion discrimination (Bartlett, et al., 2004; Dhall, Asthana, Goecke, & Gedeon, 2011; Valenti, Sebe, Gevers, & Cohen, 2008; Saeed et al., 2014), and shown how complex or blended emotions can be constructed from more basic primitives (Du et al., 2014; Young et al., 1997). The FACS, in combination with similar approaches from computer graphics (Facial Animation Parameters, FAPs: Pardas, Bonafonte, & Landabaso, 2002) provide a representational repertoire that can be used to map out the temporal trajectory of different emotions (Jack, Garrod, & Schyns, 2014; Pantic & Patras, 2006; Valstar & Pantic, 2006; Zhang & Ji, 2005), to synthesize different emotions and emotion combinations (Roesch, 2011; Mäkäräinen & Takala, 2009), and to design effective algorithms for automated emotion classification (Bettadapura, 2012; Martinez and Du, 2012; Pantic, 2009; Poria, Cambria, Hussain, & Huang, 2015). Together, this research shows that it is possible to characterize the motion parameters and other perceptual invariants that support extracting emotion from dynamic facial expressions. A challenge for the present thesis is to generate a
similarly explicit account of the conceptual features or regularities that allow for inferring emotions from their eliciting situations.

### 0.2.2 The development of facial emotion recognition

Sensitivity to others’ emotional expressions is evident early in life. By the second half of the first year, infants reliably discriminate between different emotional expressions (particularly for coarse distinctions such as positive versus negative affect), and exhibit emotion-specific neural signatures and behavioral responses (Leppänen & Nelson, 2006; Nelson, 1987; Tronick, 1989). Like adults, infants show categorical emotion discrimination that transcends physical similarity of the stimuli—differentiating between expression morphs that cross an emotion-category boundary, but not expressions that are comparably dissimilar but fall within an emotion category (Kotsoni, de Haan, & Johnson, 2001; Leppänen, Richmond, Vogel-Farley, Moulson, & Nelson, 2009)—and they discriminate emotional expressions in a way that generalizes across individual identity and is specific to emotionally relevant stimuli (i.e. upright faces: Kestenbaum & Nelson, 1990). By this age, infants are also able to detect congruency between facial and vocal expressions (Grossmann, Striano, & Friederici, 2006; Vaillant-Molina, Bahrick, & Flom, 2013; Walker-Andrews, 1997; Walker-Andrews & Grolnick, 1983; see Kahana-Kalman & Walker-Andrews, 2001 for possible evidence at 3.5 months). Moreover, there is evidence that infants not only perceptually categorize different facial expressions, but also extract meaningful information from them. By seven months, infants look to caregivers’ faces for emotional information in ambiguous situations (Striano & Vaish, 2006), and as early as 12 months, infants use an adult’s emotional expressions to guide their own behavior (Hertenstein & Campos, 2004; Moses, Baldwin, Rosicky, & Tidball, 2001; Mumme et al., 1996; Mumme & Fernald, 2003).

Certain emotion sensitivities are evident even prior to 6 months, though the results at younger ages are more varied. Emotion discrimination has been reported in infants as young as 3-5 months in studies using preferential looking paradigms (Haviland & Lelwica, 1987; Kuchuk, Vibbert, & Bornstein, 1986; LaBarbera, Izard, Vietze, & Parisi, 1976; Schwartz, Izard, & Ansul, 1985), habituation paradigms (Barrera & Maurer, 1981; Serrano, Iglesias, & Loeches, 1995; Young-Browne, Rosenfeld, & Horowitz, 1977) and peekaboo and startle
methods (Balaban, 1995; Montague & Walker-Andrews, 2001), and young infants also exhibit emotive behaviors (smiling, leaning towards the screen, avoidance movements) that are matched to different emotional expressions (Serrano et al., 1995; Walker-Andrews, 1998). In fact, several studies report discrimination of different emotional expressions even in neonatal infants, as measured by both ERP and behavioral measures (Cheng, Lee, Chen, Wang, & Decety, 2012; Farroni, Menon, Rigato, & Johnson, 2007; Field, Woodson, Greenberg, & Cohen, 1982; Zhang et al., 2014). This early-emerging sensitivity to emotional expressions, combined with observed heritability in facial emotion perception (Anokhin, Golosheykin, & Heath, 2010; de Achával et al., 2010; see also Grossmann et al., 2011), provides support for the proposal that humans co-evolved emotion production and recognition abilities, with perceptual mechanisms for recognizing specific emotional displays being part of an innate perceptual repertoire (Darwin, 1872; Ekman & Friesen, 1971; Fridlund, 2014).

Together, these experiments with adults and infants provide a growing understanding of the mechanisms that support facial emotion perception, and their emergence over human development. Although faces have been most extensively studied compared to perceptual displays in other modalities, similarly bodies of research are emerging on the perception of emotion from vocal expressions and body posture/gait, in both adults (auditory: Bachorowski, 1999; Bachorowski & Owren, 2003; Juslin & Laukka, 2003; Scherer, 2003; bodily: App, McIntosh, Reed, & Hertenstein, 2011; Atkinson, 2009; Atkinson et al., 2004; Clarke, Bradshaw, Field, Hampson, & Rose, 2005; de Gelder, de Borst, & Watson, 2014) and infants (auditory: Fernald, 1993; Flom & Bahrick, 2007; Grossmann, 2010a; Grossmann, Striano, & Friederici, 2005; Walker-Andrews & Grolnick, 1983; Walker-Andrews & Lennon, 1991; bodily: Missana, Rajhans, Atkinson, & Grossmann, 2014; Zieber, Kangas, Hock, & Bhatt, 2013, 2014).

---

1 Although a number of studies report that facial emotion recognition improves with age (Chronaki, Hadwin, Garner, Maurage, & Sonuga-Barke, 2014; Montirosso, Peverelli, Frigerio, Crespi, & Borgatti, 2010; Tonks, Williams, Frampton, Yates, & Slater, 2007; Tottenham, Hare, & Casey, 2011), many of these studies involve verbal labeling and other task demands that could mask the recognition abilities of younger children. Young children exhibit adult-like perceptual signatures when recognizing facial emotion (e.g. inversion and composite effects: Durand, Gallay, Seigneuric, Robichon, & Baudouin, 2007), and the most reliable age-related changes seem to be specific to distinguishing particular subsets of emotions, such as anger and disgust (Gao & Maurer, 2010; Herba, Landau, Russell, Ecker, & Phillips, 2006; Kolb, Wilson, & Taylor, 1992; Widen & Russell, 2013).
0.2.3 Neural mechanisms for emotion perception

As with behavioral research, studies investigating the neural mechanisms underlying emotion attribution have generally focused on the perception of overt emotional displays, particularly facial expressions (Adolphs, Damasio, Tranel, & Damasio, 1996; Haxby, Hoffman, & Gobbini, 2000; Phillips, Drevets, Rauch, & Lane, 2003; Vuilleumier & Pourtois, 2007). Motivated by basic emotion theory, initial studies suggested that there might be distinct neural circuits associated with different emotional states (e.g. amygdala for fear: Adolphs, Tranel, Damasio, & Damasio, 1995; Davis, 1992; insula for disgust: Phillips et al., 1997; Wicker et al., 2003; orbitofrontal cortex for anger: Blair et al., 1999), and that recognition of different emotions might recruit these emotion-specific neural mechanisms (Fusar-Poli et al., 2009; Phan, Wager, Taylor, & Liberzon, 2002; Vytal & Hamann, 2010). However, at present, there is little evidence for consistent and selective associations between discrete brain regions and specific emotions (Kober et al., 2008; Lindquist, Wager, Kober, Bliss-Moreau, & Barrett, 2012). For example, the amygdala—a purported "fear" mechanism—responds to a range of arousing or emotionally salient stimuli that do not involve fear (e.g. Ewbank, Barnard, Croucher, Ramponi, & Calder, 2009; Jenison, Rangel, Oya, Kawasaki, & Howard, 2011; Weierich, Wright, Negreira, Dickerson, & Barrett, 2010). Others have suggested that there may be a small number of core dimensions (e.g. valence and arousal) that are implemented in distributed neural systems (e.g. limbic regions such as anterior insula, anterior cingulate cortex, and amygdala) and recruited across a range of different emotions (Barrett & Bliss-Moreau, 2009; Kober et al., 2008; Lindquist, Satpute, Wager, Weber, & Barrett, 2015; Touroutoglou, Lindquist, Dickerson, & Barrett, 2015; Wilson-Mendenhall, Barrett, & Barsalou, 2013).

However, while emotion-specific representations could, in principle, take the form of a consistent response across voxels in different regions (e.g. the amygdala representing fear), a fruitful approach in other domains of perception and cognition has been to characterize neural representations at a spatial scale smaller than that of entire regions (e.g. representations internal to face-selective regions: Anzellotti, Fairhall, & Caramazza, 2013; Freiwald & Tsao, 2010a; Ramírez, Cichy, Allefeld, & Haynes, 2014). With this approach, we can understand the computational properties of a given brain region by characterizing the features or dimensions that
are distinguished in its neural patterns, and the dimensions to which those neural distinctions are invariant (DiCarlo & Cox, 2007; Kriegeskorte & Kievit, 2013; Kriegeskorte & Kreiman, 2012). This idea of a “population code”—a population of neurons that can be thought of as representing or “coding for” the dimensions or distinctions that can be read out (“decoded”) from their distributed pattern of activity—has proved extremely useful for understanding the neural representations involved in domains of visual perception (e.g. characterizing the series of transformations that make up object recognition in IT: DiCarlo et al., 2012; Khaligh-Razavi & Kriegeskorte, 2014; Kiani, Esteky, Mirpour, & Tanaka, 2007; Yamins et al., 2014). While functional MRI does not allow for directly measuring the information in a population of neurons, neuroimaging research has adopted a comparable approach (employed throughout this thesis) of analyzing the information contained across distributed patterns of voxels (Haxby, Connolly, & Guntupalli, 2014; Kriegeskorte, 2009; Kriegeskorte, Goebel, & Bandettini, 2006; Norman, Polyn, Detre, & Haxby, 2006). As described below, such methods have been effectively used in a number of studies to identify regions in which the population code distinguishes a space of distinct emotions.

A second limitation to the studies reviewed above is that they do not aim to distinguish neural responses to emotional facial expressions from responses to emotionally arousing stimuli more broadly. Thus, it is unclear whether the implicated regions (e.g. amygdala, insula) are involved in the perception and attribution of emotional states, or whether they reflect emotional states induced in the participant themselves. A recent meta-analysis comparing neural responses to emotional faces and emotional scenes suggests that there may be regions that contribute specifically to the perception of emotion from faces (Sabatinelli et al., 2011), including a network of face-selective regions in the ventral stream: the occipital face area (OFA), fusiform face area (FFA) and superior temporal sulcus (STS) (see De Winter et al., 2015; Engell & Haxby, 2007; Fox, Moon, Iaria, & Barton, 2009; Ganel, Valyear, Goshen-Gottstein, & Goodale, 2005; Vuilleumier & Pourtois, 2007; Xu & Biederman, 2010).

The STS, in particular, is hypothesized to play a key role in facial emotion recognition. A large body of research implicates the STS in the processing of dynamic faces (De Winter et al., 2015; Fisher & Freiwald, 2015; Harris, Young, & Andrews, 2012; Pitcher, Dilks, Saxe, Triantafyllou, & Kanwisher, 2011; Reinl & Bartels, 2014),
suggesting that this region may encode transient, motion-based properties such as emotional expression (Andrews & Ewbank, 2004; Calder & Young, 2005; Haxby et al., 2000). Consistent with this hypothesis, recent research finds that the overall magnitude of response is higher for emotional relative to neutral faces (Engell & Haxby, 2007), and that population responses contain information about distinct emotional expressions (Baseler, Harris, Young, & Andrews, 2014; Flack et al., 2014; Furl, Hadj-Bouziane, Liu, Averbeck, & Ungerleider, 2012; Said, Moore, Engell, Todorov, & Haxby, 2010; Said, Moore, Norman, Haxby, & Todorov, 2010). Having identified candidate substrates for facial emotion recognition, exciting recent research is now beginning to characterize the specific emotion-relevant properties (e.g., parts-based facial kinematics, configural information) that structure neural representations in different parts of the face hierarchy (Deen et al., in review; Flack et al., 2014).

Emotions expressed via other modalities may be represented in other cortical regions. For example, bodily expressions of emotion activate a set of regions that only partially overlaps with those responding to emotional faces, including extrastriate body area (EBA) and parts of superior temporal sulcus (de Gelder, 2006; Kret, Roelofs, Stekelenburg, & de Gelder, 2013; Tamietto & de Gelder, 2010). Similarly, regions such as the superior temporal gyrus, middle temporal gyrus, and secondary auditory cortex respond to emotional speech and vocalizations relative to neutral speech (Ethofer et al., 2006, 2011; Wiethoff et al., 2008), and distributed patterns of activity within these regions support decoding of specific emotions (Ethofer, Van De Ville, Scherer, & Vuilleumier, 2009; see also Bestelmeyer, Mautage, Rouger, Latinus, & Belin, 2014).

0.3 THE PRESENT RESEARCH: INFERRING EMOTIONS

While this prior research sheds light on the mechanisms by which we recognize emotions from different perceptual cues, the recognition of overt displays captures only a fraction of human emotion knowledge. I argue that focusing on perceptual processing provides a limited perspective on emotion attribution abilities for several reasons. First, overt expressions are only one kind of cue that supports attribution. Given that we also use situations and causal context to reason about what others are feeling, a full understanding of emotion attribution
requires characterizing these inferential mechanisms, and their relationship to the perceptual processes studied thus far. In fact, a number of recent studies find that surrounding context can alter or dominate the perception of emotion from facial expressions (Aviezer et al., 2008; Aviezer, Dudarev, Bentin, & Hassin, 2011; Barrett, Lindquist, Bliss-Moreau, et al., 2007; Barrett, Mesquita, & Gendron, 2011; Barrett & Kensinger, 2010; Hassin, Aviezer, & Bentin, 2013).

A second reason to move beyond the study of facial emotion is that the space of emotions we represent and attribute may be much larger than that captured in overt expressions (Fontaine et al., 2007a). As discussed at length in Chapter 3, there have been a number of attempts to encode emotions in terms of some low dimensional space of affective primitives. By focusing on facial expressions—a potentially narrow and noisy window into human emotion—this research may be fundamentally limited in its ability to capture the dimensionality of human emotion concepts. To more fully characterize the content and structure of intuitive emotion knowledge, we need stimuli and tasks that have the potential to span the full range of concepts in this domain.

Finally, focusing exclusively on overt expressions may lead to a distorted view of the cognitive architecture underlying mental state reasoning more broadly (see Mitchell & Phillips, 2015). Many have emphasized the differences between emotion attribution and belief-based reasoning (“hot” vs “cold” ToM: Brothers & Ring, 1992; affective vs. epistemic perspective taking: S. G. Shamay-Tsoory, Tomer, Goldsher, Berger, & Aharon-Peretz, 2004; cognitive vs. emotional empathy: Blair, 2005), and proposed that these may be dissociable abilities implemented in distinct neural systems (e.g. Blair, 2005). For example, it has been argued that understanding others’ emotions relies on simpler mechanisms that support “simulating” affective states within a putative mirroring system (Braadbaart, de Grauw, Perrett, Waiter, & Williams, 2014; Carr, Iacoboni, Dubeau, Mazziotta, & Lenzi, 2003; Nummenmaa, Hirvonen, Parkkola, & Hietanen, 2008; Pfeifer, Iacoboni, Mazziotta, & Dapretto, 2008; Schulte-Ruther et al., 2011; van der Gaag, Minderaa, & Keysers, 2007). However, these studies investigated emotion understanding in the context of recognizing emotions from familiar perceptual cues like facial expressions. To determine the relationship between mechanisms supporting emotion understanding and
other aspects of ToM, we must also investigate the aspects of emotion attribution that are most similar to belief reasoning—namely, those that require theory-like inferences based on abstract situational information.

Thus, this thesis aims to investigate the representations that support reasoning about others’ emotions based on the situations that elicit them. A premise of this work is that humans possess intuitive knowledge of different emotions that captures their causal structure; that is, people interpret overt displays as communicating internal affective states that are caused by an agent’s goals, beliefs and attitudes towards the situations they encounter. I suggest that this knowledge forms part of a more general intuitive theory that allows us to explain and predict the mental states of others, and is at the heart of many of our most sophisticated social-cognitive capacities. This thesis provides an initial attempt to 1) articulate the precise content of intuitive emotion knowledge, 2) characterize how this knowledge is represented in the human brain, and 3) explore the origin of abstract emotion concepts in early human development.

0.3.1 Chapter 1

In Chapter 1, I aim to uncover the neural mechanisms that support abstract, conceptual representations of emotion, and to distinguish them from perceptual representations that may be tied to lower-level parameters differentiating emotional expressions in a single modality. To date, one study has gone beyond modality-specific representations, finding that neural populations in posterior temporal cortex and the medial prefrontal cortex (MPFC) discriminate emotions in a way that generalizes across facial, vocal, and bodily displays (Peelen, Atkinson, & Vuilleumier, 2010). This cross-stimulus generalization points towards supramodal a representation of emotion that cannot be explained in terms of low-level perceptual parameters.

---

2 Researchers have described pieces of this knowledge, but focused primarily on basic emotions. For example, according to Ekman, a prototypical antecedent event for sadness would be “loss of a significant other” (Ekman, 1992). In one study, subjects were asked to describe an event in which they experienced a particular basic emotion, and researchers then extracted regularities in the conditions that elicited the different emotions. For example, anger was associated with “reversal or sudden loss of power, status, or respect”, “frustrating or interruption of a goal-directed activity”, “real or threatened physical or psychological pain”, and “judgment that the situation is illegitimate, wrong, or unfair” (Shaver et al., 1987). However, this research focused primarily on first-person experience of emotions, rather than on third-party inferences, and was limited to a small number of emotional states.
Still, there are various underlying representations that could give rise to multimodal neural responses. For example, a particularly lean interpretation would be that the association between co-occurring perceptual correlates of an emotion (e.g. a crying face and voice) leads to spreading activation between the representations of relevant modality-specific features. In this case, multimodal patterns could reflect modality-specific representations that are merely co-activated based on inputs in other associated modalities (e.g. Brosch, Grandjean, Sander, & Scherer, 2009; Gerdes et al., 2013). A related possibility is that multimodal representations reflect similarities in the information exploited in different modalities, as behavioral studies have revealed commonalities in emotion-relevant parameters across vision and audition. For example, Wheatley and colleagues found that subjects associated different emotions with combinations of features such temporal frequency and dissonance/”spikiness” that were shared across visual motion and auditory stimuli (Sievers, Polansky, Casey, & Wheatley, 2013). In this case, emotional expressions could be decoded from regions that are multimodal, but that represent more basic parameters that happen to be shared across these modalities.

A final possibility is that observed cross-modal similarities reflect a representation that plays a more complete conceptual role in emotion attribution and inference. Patterns of activity in MPFC or posterior temporal cortex could encode an abstract representation that is fully invariant to lower-level regularities that support emotion identification within a specific modality (e.g. analogous to ventral stream recognition, where invariance is achieved through a hierarchy of increasingly complex perceptual feature-detectors (e.g. Riesenhuber & Poggio, 1999; Rolls & Milward, 2000; Serre et al., 2007; Wallis & Rolls, 1997). Of course, if emotion knowledge is causal and theory-like, a simple feed-forward hierarchy of filters will surely be insufficient to explain the full process of emotion attribution. Nonetheless, there may be a final stage of the processing stream at which emotion representations not only pool over lower-level perceptual schemas, but also incorporate causal information about the context or situation. It is currently unknown whether representations in regions like MPFC play such a rich conceptual role.

To distinguish between these possibilities, I use multi-voxel pattern analysis (MVPA) to test for regions that contain information about the emotional valence of facial expressions and of short animations depicting
emotion-eliciting situations. If a region contains a fully conceptual representation of emotion, its population code should be invariant to whether the emotion was perceived from a canonical expression, or inferred from an abstract eliciting cause. As described above, previous studies implicate a number of different regions in the perception of emotional expressions, including regions that may be modality specific, and others that may be multimodal\(^3\). Given both its supramodal discrimination properties (Peelen et al., 2010), and its involvement in theory-like mental state reasoning (Bruneau, Pluta, & Saxe, 2012a; Dodell-Feder, Koster-Hale, Bedny, & Saxe, 2011; Mitchell & Phillips, 2015; Saxe & Kanwisher, 2003), the MPFC is a particularly strong candidate for implementing a fully abstract, conceptual representations of attributed emotion. If the MPFC contains such a representation, a model trained to discriminate emotional valence for one kind of input (e.g. facial expressions) should be able to successfully decode the emotional valence of another kind of input (e.g. eliciting situations). By testing whether different neural populations can classify emotional states within and between stimulus types, the research in Chapter 1 can provide constraints on the possible computational roles of different regions, and identify candidate neural substrates for distinct levels of representation (modality-specific, multimodal, and conceptual).

### 0.3.2 Chapter 2

In Chapter 2, I examine the developmental origins of basic emotion concepts. A fundamental question in cognitive development concerns how flexible, theory-like knowledge, such as our intuitive knowledge of emotions, emerges during human development. What are the initial primitives available for reasoning about others’ emotions, and what processes yield the mature state of knowledge? As reviewed above, studies with infants have tended to focus on discrimination of overt expressions, an ability that is evident early and possible innately (Grossmann, 2010b). But what conceptual content is assigned to these emotional displays early in development? One possibility is that infants possess perceptual schemas that allow them to identify different emotional

---

\(^3\) Superior temporal cortex in particular is argued to play a role in multisensory integration of affective information (Calvert, 2001; Hagan et al., 2009; Hagan, Woods, Johnson, Green, & Young, 2013; Kreifelts, Ethofer, Grodd, Erb, & Wildgruber, 2007; Kreifelts, Ethofer, Shiozawa, Grodd, & Wildgruber, 2009; Park et al., 2010; Robins, Hunyadi, & Schultz, 2009; Stevenson & James, 2009).
expressions (perhaps innately triggering similar affective states in the infant) but that they lack a conceptual representation of these displays as related to emotion inducing events. In fact, some have argued that emotion concepts depend critically on natural language (Lindquist & Gendron, 2013; Lindquist, MacCormack, & Shablack, 2015) and that even perceptual discrimination of different emotions should be weak the absence of language (Barrett, Lindquist, & Gendron, 2007; Gendron, Lindquist, Barsalou, & Barrett, 2012; Lindquist, Barrett, Bliss-Moreau, & Russell, 2006). Alternatively, infants could be equipped with representations of emotions that play a relatively rich theoretical role, where different expressions are interpreted in relation to the events that elicit them, and the subsequent behaviors they predict.

Chapter 2 tests whether preverbal infants have any understanding of the context that elicits different emotional reactions, specifically focusing on the relationship between positive and negative emotions and the outcomes of simple goals. A large body of research demonstrates that infants identify and reason about the goal-directed actions of others (Csibra & Gergely, 2007; Woodward, 1998); if infants’ goal representations play a conceptual role similar to that of adults, they should understand how goal outcomes relate to affective states and subsequent displays. Pieces of this knowledge have been demonstrated by early childhood; in fact, studies of specific emotion concepts in children provide some of the most thorough research on intuitive theories of emotion conducted to date (Nunner-Winkler & Sodian, 1988; Saarni & Harris, 1991; Henry M. Wellman & Banerjee, 1991). Studies of children’s “emotion scripts”, for example, have aimed to characterize emotional development in terms of the construction of emotion-specific sequences that associate emotion labels with a set of causes, consequences, and overt expressions (Russell, 1990; Saarni & Harris, 1991; Widen, Pochedly, & Russell, 2015; Widen & Russell, 2004, 2010). By ages 7-8, children can infer that a person will feel emotions such as disgust or embarrassment in response to prototypical scenarios, and emotions like anger, sadness, and fear are attributed even 3-4 years earlier (Widen & Russell, 2010). Of course, by these ages, children have had years of
verbal communication about emotional states, including explicit instruction about the emotional consequences of different types of events (Brownell, 2013; Denham & Kochanoff, 2002; Dunn, Brown, & Beardsall, 1991)\(^4\).

Do even preverbal infants also have a coherent representation of the relation between goals, outcomes, and emotions? In several experiments, I use animations in which a simple geometric character demonstrates a goal, successfully achieves or fails to achieve the goal, and reacts with positive or negative affect. If infants understand that goal outcomes yield particular emotions in others, they should have expectations about the affective responses that will be elicited by different successful and unsuccessful actions.

0.3.3 Chapter 3

In Chapters 1 and 2, I focus on reasoning about causes relating to a single emotional distinction: differentiating positive versus negative affect. However, humans possess a rich vocabulary of emotional states that extends far beyond this binary distinction. Thus, I build on these first two studies by investigating a more nuanced space of emotion concepts. If people possess theory-like knowledge of the eliciting situations for a wide range of different emotions, they should be able to make fine-grained attributions from descriptions of antecedent events. In Chapter 3, I show that subjects indeed make consistent attributions of the emotions others feel in a range of situations, and I identify a set of abstract properties of these situations (i.e. an “event” feature space) that well explains the emotional discriminations subjects make behaviorally. I then test whether this feature space can capture the similarity space of neural patterns in the brain regions identified in Chapter 1. Specifically, I measure BOLD responses to verbal vignettes describing events that would elicit one of 20 distinct emotions, and use a combination of multi-voxel pattern analysis (MVPA) and Representational Similarity Analysis (RSA) to identify brain regions containing abstract and high-dimensional representations of attributed emotion.

\(^4\) More, while this kind of research shows that children have the ability to infer emotions from some situations, how they do so remains less clear, as the stimuli tend to involve a small number of highly stereotypical eliciting events (“It was Joan’s birthday. All her friends came to her birthday party and gave her presents; How did Joan feel?”: Widen & Russell, 2010). Thus, this research generally leaves open whether children have an understanding of abstract properties of the events that elicit different emotions, or are simply familiar with the emotions elicited in specific situations. One exception to this is research on children’s inferences about emotions based on desires. By 3 years old, children can flexibly incorporate information about an individual’s desires and intentions to reason about how the person will feel in a given situation (Russell, 1990; Henry M. Wellman & Banerjee, 1991; Yuill, 1984).
As mentioned above, existing theories have aimed to reduce the space of human emotion into some combination of simple dimensions or categories (e.g. valence and arousal, or a set of 4-6 basic emotions). Similarly, prior literature on the neural basis of emotion recognition has focused on very simple affective discriminations (e.g. positive vs. negative). In Chapter 3, I test whether fine-grained emotional distinctions (e.g. someone feeling proud vs. grateful) can be decoded from brain regions selective for reasoning about mental states. I then examine the representational structure within these regions, testing whether the observed neural similarity structure is well captured by a high dimensional “event” feature space, or whether it may be reducible to more primitive affective dimensions such as valence and arousal.

0.4 CONCLUSIONS

In sum, this thesis aims to characterize the representations that underpin human emotion knowledge and inference, including how the relevant concepts develop, and how this information is encoded in the human brain. I argue that characterizing the mechanisms that support emotion perception and attribution will require studying emotions as more than mere bodily and vocal displays. The research described in the following three chapters demonstrates that humans reliably infer others’ emotions from antecedent causes, and aims to characterize the intuitive theory that makes these inferences possible. In Chapter 2, I demonstrate that even early in development, emotional expressions are analyzed in relation to preceding events. Moreover, I show that abstract properties of eliciting events not only explain subjects’ attributions of emotions to particular situations, but also structure neural representations of emotions in regions involved in mental state reasoning. Together, Chapters 1 and 3 suggest that regions such as MPFC link perceptual and inferential mechanisms to form abstract representations of emotional states, and that neural populations within this and other brain regions encode a rich space of emotions that cannot be reduced to a small set of affective primitives. These findings show that it is possible to characterize the fine-grained representational structure of a high-level, uniquely human reasoning capacity, and constitute an initial a step towards understanding how those representations emerge within cortical hierarchies, and over human development.
A common neural code for perceived and inferred emotion

1.0 ABSTRACT

Although the emotions of other people can often be perceived from overt reactions (e.g. facial or vocal expressions), they can also be inferred from situational information in the absence of observable expressions. How does the brain make use of these diverse forms of evidence to generate a common representation of a target’s emotional state? In the present research, we identify neural patterns that correspond to emotions inferred from contextual information, and find that these patterns generalize across different cues from which an emotion can be attributed. Specifically, we use functional MRI to measure neural responses to dynamic facial expressions with positive and negative valence, and to short animations in which the valence of a character’s emotion could be identified only from the situation. Using multi-voxel pattern analysis, we test for regions that contain information about a target’s emotional state, identifying representations specific to a single stimulus type, and representations that generalize across stimulus types. In regions of medial prefrontal cortex (MPFC), a classifier trained to discriminate emotional valence for one stimulus (e.g. animated situations) could successfully discriminate valence for the remaining stimulus (e.g. facial expressions), indicating a representation of valence that abstracts away from perceptual features and generalizes across different forms of evidence. Moreover, in a subregion of MPFC, this neural representation generalized to trials involving subjectively experienced emotional events, suggesting partial overlap in neural responses to attributed and experienced emotions. These data provide a step towards understanding how the brain transforms stimulus-bound inputs into abstract representations of emotion.
1.1 INTRODUCTION

To recognize someone’s emotion, we can rely on facial expression, tone of voice, even body posture. Perceiving emotions from these overt expressions poses a version of the "invariance problem" faced across perceptual domains (see Ullman, 1998; Dicarlo et al., 2012): we recognize emotions despite variation both within modality (e.g. sad face across viewpoint and identity) and across modalities (e.g. sadness from facial and vocal expressions). Emotion recognition may therefore rely on bottom-up extraction of invariants within a hierarchy of increasingly complex feature-detectors (e.g. Tanaka, 1993). However, we can also infer emotions in the absence of overt expressions by reasoning about the situation a person encounters (Ortony et al., 1990, Zaki et al., 2008; Scherer & Meuleman, 2013). To do so, we rely on abstract causal principles (e.g. social rejection causes sadness) rather than direct perceptual cues. Ultimately, the brain must integrate these diverse sources of information into a common code that supports empathic responses and flexible emotion-based inference.

What neural mechanisms underlie these different aspects of emotion recognition? Previous neuroimaging studies have revealed regions containing information about emotions in overt expressions: different facial expressions, for example, elicit distinct patterns of neural activity in the superior temporal sulcus and fusiform gyrus (Said et al., 2010a,b; Harry et al., 2013; see also Pitcher, 2014). In these studies, emotional stimuli were presented in a single modality, leaving it unclear the precise dimensions represented in these regions. Given that facial expressions can be distinguished based on features specific to the visual modality (e.g. mouth motion, eyebrow deflection, eye aperture; Ekman & Rosenberg, 1997; Oosterhof & Todorov, 2009), face-responsive visual regions could distinguish emotional expressions based on such lower-level features.

To represent what is in common across sad faces and voices, the brain may also compute multimodal representations. In a recent study (Peelen et al. 2010), subjects were presented with overt facial, bodily, and vocal expressions: in posterior temporal cortex (lpSTC) and middle medial prefrontal cortex (MMPFC), the pattern of response across different modalities was more similar for the same emotion than for different emotions. Thus, emotional stimuli sharing no low-level perceptual features seem to be represented similarly in these regions.

However, we not only recognize emotions from canonical perceptual cues, but also infer emotions from
causal context alone. We identify emotions in the absence of familiar expressions, even for situations we have never observed or experienced. In the present study, we test for neural representations of emotional valence that generalize across both overt facial expressions and emotions inferred from the situation a character is in. We first identify neural patterns that contain information about emotional valence for each type of stimulus. We then test whether these neural patterns generalize across the two stimulus types, the signature of a common code integrating these very different types of emotional information. Finally, we investigate whether attributing emotional experiences to others and experiencing one’s own emotions recruit a common neural representation by testing whether these same neural patterns generalize to emotional events experienced by participants themselves.

### 1.2 MATERIALS AND METHODS

#### 1.2.a Summary

In Experiment 1, we used functional magnetic resonance imaging (fMRI) to measure blood-oxygen-level dependent (BOLD) responses to emotional facial expressions and to animations depicting a character in an emotion-eliciting situation. While emotion-specific representations could, in principle, take the form of a uniform response across voxels in a region (detectable with univariate analyses), prior research has yielded little evidence for consistent and selective associations between discrete brain regions and specific emotions (Fusar-Poli et al., 2009; Lindquist et al., 2012). Thus, the present research uses multivariate analyses that exploit reliable signal across distributed patterns of voxels to uncover neural representations at a spatial scale smaller than that of entire regions (Haxby et al., 2001; Kamitani & Tong, 2005; Kriegeskorte et al., 2006; Norman et al., 2006). With this approach, we test for representations of emotional valence that are specific to a particular type of stimulus (facial expressions or causal situations) and representations that generalize across the two stimulus types. To identify stimulus-independent representations, we trained a pattern classification algorithm to discriminate emotional valence for one stimulus type (e.g. dynamic facial expressions) and tested its ability to discriminate valence for the remaining type (e.g. animations depicting causal situations). Thus, for each region of interest, we test whether there is a reliable neural pattern that supports classifying emotions when trained and tested on facial expressions,
when trained and tested on situations, and when requiring generalization across facial expressions and situations.

We then test whether attributing emotions to others engages neural mechanisms involved in first-person experience of emotion. Previous research has implicated MPFC not only in emotion attribution, but also in subjective experience of emotional or rewarding outcomes (Lin et al., 2012, Winecoff et al., 2013; Clithero & Rangel, 2013; Chikazoe et al., 2014). However, the relationship between experienced reward and emotion attribution remains poorly understood. In Experiment 2, we measured BOLD responses to positive and negative situations for another individual (replicating Experiment 1) and to trials in which subjects themselves experienced positive and negative outcomes (winning and losing money). Again, we test whether there is a reliable neural pattern that supports classifying the valence of events when trained and tested on third-party situations, when trained and tested on first-person rewards, and when requiring generalization across third-person and first-person experiences.

1.2.b Regions of interest

Based on prior literature (Peelen et al., 2010), our regions of interest for abstract, conceptual representations of emotion were the pSTC and MMPFC. We localized in individual subjects a middle MPFC ROI comparable to Peelen et al. (2010), using a standard social vs. nonsocial contrast (Saxe & Kanwisher, 2003; Dodell-Feder et al., 2011; see below). Because pSTC could not be identified by standard localizer tasks, we identified bilateral group ROIs based on the peak coordinate from Peelen et al (2010). Our primary analyses target these three ROIs, accounting for multiple comparisons with a corrected alpha of 0.05/3 (0.017).

In addition to the MMPFC region identified by Peelen and colleagues (2010), adjacent regions of dorsal and ventral MPFC have been strongly implicated in studies of emotion and affective value (Amodio & Frith, 2006; Hynes et al., 2006; Völlm et al., 2006; Etkin et al. 2011). Moreover, the MPFC is part of a larger set of regions—the posterior cingulate/precuneus (PC), bilateral temporal parietal junction (TPJ) and right anterior temporal lobe (rATL)—that are reliably recruited when reasoning about others’ mental states (Saxe & Kanwisher, 2003, Mitchell, 2009), including emotional states (Zaki et al., 2010; Spunt & Liberman, 2012; Bruneau et al., 2012). This set of 6 regions—DMPFC, VMPFC, rTPJ, lTPJ, PC, and rATL (in addition to
MMPFC described above)—were identified in individual subjects using the social vs. nonsocial contrast (described below). We test these remaining regions for representations of both perceived and inferred emotions (with an alpha of 0.05/6 (0.008) to correct for comparisons across these 6 ROIs).

To test for modality-specific representations, we localized regions that might contain information specific to overt facial expressions: the rmSTS, hypothesized to code for facial motion parameters (Carlin et al., 2011; Calder et al., 2007; Pelphrey et al., 2005), and face-selective patches in right occipitotemporal cortex thought to code for identity-relevant face features (occupital face area: rOFA, fusiform face area: rFFA; Kanwisher & Yovel, 2006). For this analysis, we again correct for multiple comparisons using an alpha of 0.017 (0.05/3).

Finally, in experiment 2, we examined how the mechanisms involved in third-person attribution of emotional states relate to mechanisms involved in processing first-person subjective value. To do so, we identified a region of OFC/VMPFC that has been previously implicated in processing reward/emotional value (Plassmann et al., 2007; Kable & Glimcher, 2007; Chib et al., 2009; Winecoff et al., 2013; Chikazoe et al., 2014). We used a mask derived from two recent meta-analyses (Clithero & Rangel, 2013; Barta et al., 2013) to investigate neural responses in an anatomical region of OFC/VMPFC in which neural responses have been shown to consistently correlate with reward value across reward types and decision contexts. Note that this mask is only partially overlapping with the searchspace used to identify VMPFC responses to theory of mind (in Experiment 1).

1.2.c Participants

21 right-handed participants (20-43 years, M_{age}= 26.84, 14 male) were recruited for Experiment 1. 16 right-handed participants (19-40 years, M_{age}= 27.88, 7 male) were recruited for Experiment 2. All participants had normal or corrected-to-normal vision and no history of neurological or psychiatric disorders, and gave written, informed consent in accordance with the requirements of the MIT institutional review board.

1.2.d fMRI tasks and stimuli

In Experiment 1, each subject participated in several behavioral tasks as well as three fMRI tasks: an Emotion Attribution task, and two tasks used to localize regions involved in theory of mind and face perception. Subjects in Experiment 2 completed only the Emotion Attribution task and the theory of mind localizer.
**Emotion Attribution task:** In the Emotion Attribution task (see Figure 1.1), subjects viewed brief video clips designed to elicit the attribution of an emotional state to a target (figure depicts static photos similar to video clips used in the study). The task consisted of video clips of people expressing a positive (happy/smiling) or negative (sad/frowning) emotion (*expressions condition*), and brief animations in which a simple geometric character experienced an event that would elicit positive or negative emotion (*situations condition*). In the situations condition, no emotion was expressed, but the character’s emotional state could be inferred based on the character’s goals and the event outcome. To ensure consistent attributions of emotional valence, independent subjects on Amazon Mechanical Turk (N=16 per item) rated the stimuli from 1-7 (negative to positive valence): \( M(\text{SEM})_{\text{pos-faces}}=5.597(0.077); M(\text{SEM})_{\text{neg-faces}}=2.694(0.084); M(\text{SEM})_{\text{pos-situations}}=5.401(0.068); M(\text{SEM})_{\text{neg-situations}}=2.695(0.058) \). Each stimulus type was further divided into 2 subcategories: male and female for facial expression clips, and social and nonsocial for situation clips. In the nonsocial condition, the character demonstrated an instrumental goal, and then achieved or failed to achieve it (e.g. attempted to climb a hill and succeeded or tumbled to the bottom); in the social condition, there were multiple agents who acted prosocially or antisocially to the target character (e.g. included or excluded the target from their group). This yielded a total of 8 stimulus conditions (male-positive (MP), male-negative (MN), female-positive (FP), female-negative (FN), social-positive (SP), social-negative (SN), nonsocial-positive (NP), nonsocial-negative (NN). Because the face stimuli involved a close-perspective view on a single entity, these stimuli were presented at 7.8 x 7.4° visual angle, while the context animations were presented at 16.7x12.5°. We used dynamic, naturalistic facial expressions from movies, which are relatively uncontrolled compared to artificial stimuli (e.g. face morphs). However, our main interest is in representations that generalize to animations in the situations condition; low-level visual confounds that generalize across the two perceptually distinct stimulus sets are therefore highly unlikely. An advantage of these stimuli in the present design is that they achieve an unusual a balance between external validity (Zaki & Ochsner, 2009; Spunt & Liberman, 2012) and experimental control.

The experiment consisted of 8 runs (9.43min/run), each containing 6 stimuli in each of the 8 conditions, for a total of 48 stimuli per condition. Each condition contained 24 semantically distinct events, each of which was
presented twice over the course of the experiment with superficial transformations (the background scene for context animations and a minor luminance change for facial expressions) and the left-right orientation varied across the two presentations. Each clip was presented at fixation for 4 seconds, followed by a 1750ms window during which subjects made a behavioral response, and a 250ms blank screen. Subjects were instructed to press a button to indicate the intensity of the character’s emotion in each event (1 to 4, neutral to extreme), which focused subjects’ attention on the character’s emotional state, but ensured that motor responses (intensity) were orthogonal to the discrimination of interest (valence). The clips were presented in a jittered, event-related design and a central fixation cross was presented between trials with a variable inter-stimulus interval of 0-14 seconds. Optseq2 was used to create efficient stimulus presentation schedules with a first-order counterbalancing constraint such that each condition preceded each other with approximately equal probability across the experiment. Assignment of conditions to positions within this sequence were randomized across participants. The order of individual stimulus clips for a given condition was chosen pseudo-randomly for each participant, with the constraint that repetitions of each stimulus occurred in the same even-odd folds as the first presentation (e.g. an event presented in run 2 would be repeated in run 6, and an event presented in run 3 would be repeated in run 7).

**Figure 1.1.** Task structure for Emotion Attribution task. Events: 4s clip, 2s response. Stimuli included two stimulus types (situation stimuli and facial expression stimuli) and two valence categories (positive and negative valence).
In Experiment 2, subjects completed a modified and abbreviated version of this task (4 runs). On 50% of trials, subjects viewed nonsocial situation stimuli from Experiment 1 (96 total trials); on the remaining trials, subjects were presented with positive and negative events in which they either gained or lost money from a post-scan bonus (reward condition, see Figure 1.2). On each reward trial, subjects viewed a cycle of 20 rapidly presented random monetary values (2 seconds total), followed by the reward outcome for the trial, shown in green (2 seconds). Negative values ranged from -$0.20 to -$1.00 and positive values ranged from +$0.20 to +$2.00; this asymmetry allowed subjects to have net gain for their bonus, and accounted for the fact that losses are experienced more strongly than comparable gains (Tversky & Kahneman, 1991). The experimental design and behavioral task were identical to Experiment 1, except that subjects were asked to rate the character’s emotional intensity on the situation trials and their own emotional intensity on the reward trials.

Figure 1.2. Task structure for Experiment 2. Events: 4s trial, 2s response. Stimuli included two stimulus types (situation stimuli and reward stimuli) and two valence categories (positive and negative valence). Reward trials involved 2s of rapid cycling through random values, followed by 2 seconds during which the reward outcome was displayed.
**Theory of mind localizer:** Subjects were presented with short textual scenarios that required inferences about mental state representations (Belief condition) or physical representations such as a map, photo or painting (Photo condition) (Dodell-Feder et al., 2011; stimuli are available at http://saxelab.mit.edu/superloc.php). These two types of scenarios were similar in their meta-representational demands and logical complexity, but only the scenarios in the Belief condition required building a representation of another person's thoughts and beliefs. Scenarios were displayed for 10s, followed immediately by a true or false question (4s) about either the representation (Belief or Photo) or the reality of the situation. Each run (4.53min) consisted of 10 trials separated by 12s inter-stimulus intervals, and 12s blocks of fixation were included at the beginning and end of each run. 1-2 runs were presented to each participant. The order of stimulus type (Belief or Photo) and correct answer (True or False) were counterbalanced within and across runs.

**Face perception localizer:** Subjects viewed 2 conditions designed to identify face-selective regions: dynamic faces (video clips of human children's faces) and dynamic objects (video clips of objects in motion; from Pitcher et al., 2011). For each of these conditions, there were a total of 30 clips (3s each, separated by 333ms of blank screen), and six clips were presented in each block. This localizer also included two other conditions, biological motion and structure from motion, which were not of interest for the present analyses. All conditions were presented as 20s blocks followed by 2s of rest, and 12s blocks of fixation were included at the beginning and end of each run, as well as once in the middle of the run. Each condition was presented twice per run and subjects received 2 runs lasting 5 minutes each, with condition order counterbalanced within and across runs, and across participants. To maintain attention, subjects were required to complete a 1-back task during viewing. 2 of 21 subjects did not complete this localizer due to insufficient scan time.

**Behavioral tasks:** The AQ (Baron-Cohen et al., 2001) and the Interpersonal Reactivity Index (IRI, Davis, 1983) were completed via online Qualtrics surveys. Participants also completed an Empathic Accuracy task based on Zaki et al. (2008) and the verbal reasoning, matrices, and riddles components of the KBIT2 (Kaufman, 1990).
1.2.e Acquisition

Data were acquired on a 3T Siemens Tim Trio scanner in the Athinoula A. Martinsos Imaging Center at the McGovern Institute for Brain Research at MIT, using a Siemens 32-channel phased array head coil. We collected a high-resolution (1mm isotropic) T-1 weighted MPRAGE anatomical scan, followed by functional images acquired with a gradient-echo EPI sequence sensitive to blood-oxygen-dependent (BOLD) contrast (repetition time (TR) = 2s, echo time (TE) = 30ms, flip angle = 90°, voxel-size 3x3x3mm, matrix 64x64, 32 axial slices). Slices were aligned with the anterior/posterior commissure and provided whole-brain coverage (excluding the cerebellum).

1.2.f Analysis

Pilot data: In addition to the 21 subjects reported, 8 independent pilot subjects were analyzed to fix the parameters of the analyses reported below (e.g. size of smoothing kernel, type of classifier, method for feature selection). A general concern with fMRI analyses, and with application of machine learning techniques to fMRI data in particular, is that the space of possible and reasonable analyses is large and can yield qualitatively different results. Analysis decisions should be made independent of the comparisons or tests of interest, otherwise one risks overfitting the analysis to the data (Simmons et al., 2011). One way to optimize an analysis without overfitting is to separate subjects into an exploratory or pilot set and a validation or test set. Thus, the analysis stream reported here was selected based on the parameters that appeared to yield the most sensitive analysis of 8 pilot subjects.

Preprocessing: MRI data were preprocessed using SPM8 (http://www.fil.ion.ucl.ac.uk/spm/software/spm8/), freesurfer (http://surfer.nmr.mgh.harvard.edu/), and in-house code. Freesurfer’s skull-stripping software was used for brain extraction. SPM was used to motion correct each subject’s data via rigid rotation and translation about the 6 orthogonal axes of motion, to register the functional data to the subject’s high-resolution anatomical image, and to normalize the data onto a common brain space (Montreal Neurological Institute). In addition to the smoothing imposed by normalization, functional images were smoothed using a Gaussian filter (FWHM=5mm).

Defining regions of interest: To define individual ROIs, we used hypothesis spaces derived from
random effects analyses of previous studies (theory of mind (Dufour et al., 2013): bilateral TPJ, rATL, PC, dorsal, middle, and ventral subregions of MPFC (DMPFC, MMPFC, VMPFC); face perception (Julian et al., 2012): rSTS, rFFA, rOFA), combined with individual subject activations for the localizer tasks. The theory of mind task was modeled as a 14s boxcar (the full length of the story and question period, shifted by 1 TR to account for lag in reading, comprehension, and processing of comprehended text) convolved with a standard hemodynamic response function (HRF). A general linear model (GLM) was implemented in SPM8 to estimate beta values for Belief trials and Photo trials. We conducted high-pass filtering at 128hz, normalized the global mean signal, and included nuisance covariates to remove effects of run. The face perception task was modeled as a 22s boxcar and beta values were similarly estimated for each of condition (dynamic faces, dynamic objects, biological motion, structure from motion). For each subject, we used a one-sample t-test implemented in SPM8 to generate a map of t values for the relevant contrast (Belief>Photo for the theory of mind ROIs, Faces>Objects for the face perception ROIs), and for each ROI, identified the peak t value within the hypothesis space. An individual subject's ROI was defined as the cluster of contiguous suprathreshold voxels (minimum k=10) within a 9mm sphere surrounding this peak. If no cluster was found at p<0.001, we repeated this procedure at p<0.01 and p<.05. We masked each ROI by its hypothesis space—defined to be mutually exclusive—such that there was no overlap in the voxels contained in each functionally defined ROI. An ROI for a given subject was required to have at least 20 voxels to be included in multivariate analyses. For the pSTC region (Peelen et al, 2010), we generated a group ROI defined as a 9mm sphere around the peak coordinate from that study, as well as an analogous ROI for the right hemisphere.

**Multivariate analyses (Figure 1.3):** Multi-voxel pattern analysis (MVPA) was conducted using in-house code developed in Python using the publicly available PyMVPA toolbox (http://www.pymvpa.org/). We conducted MVPA within ROIs that were functionally defined based on individual subject localizer scans. High-pass filtering (128 Hz) was conducted on each run, and linear detrending was performed across the whole timecourse. A timepoint was excluded if it was a global intensity outlier (greater than 3 SD above the mean intensity) or corresponded to a large movement (greater than 2 mm scan-to-scan). The data were temporally
compressed to generate one voxel-wise summary for each individual trial, and these single trial summaries were used for both training and testing. Individual trial patterns were calculated by averaging the preprocessed bold images for the 6s duration of the trial, offset by 4 seconds to account for HRF lag. Rest timepoints were removed and the trial summaries were concatenated into one experimental vector in which each value was a trial's average response. The pattern for each trial was then z-scored relative to the mean across all trial responses in that voxel.

Figure 1.3. MVPA analysis procedure. Top. Valence-labeled voxel patterns (from a single ROI) used to train a linear SVM. Middle. Learned voxel-weights used to predict valence of unlabeled test data (voxel patterns not used for training). Bottom: Cross-validation schemes for testing for stimulus-specific and stimulus-independent emotion representations.
Given the high dimensionality of fMRI data, and the relatively small number of training examples available, feature selection is often useful to extract voxels likely to be informative for classification (de Martino et al., 2008; Mitchell et al., 2004; Pereira et al., 2009). Within each ROI, we conducted voxel-wise analyses of variance (ANOVAs) to identify voxels that were modulated by the task (based on the F-statistic for task vs. rest contrast). This univariate selection procedure tends to eliminate high variance, noisy voxels (Mitchell et al., 2004). Because this selection procedure is orthogonal to all of the classifications reported in the paper, it could be performed once over the whole dataset without constituting peeking, meaning that the same voxels could be used as features in each cross-validation fold. The top 80 most active voxels within the ROI were used for classification (selecting a fixed number of voxels also helps to minimize differences in the number of voxels across regions and subjects).

The data were classified using a support vector machine implemented with libSVM (http://www.csie.ntu.edu.tw/~cjlin/libsvm/; Chang & Lin, 2011). This classifier uses condition-labeled training data to learn a weight for each voxel, and subsequent stimuli (validation data not used for model training) can then be assigned to one of two classes based on a weighted linear combination of the response in each voxel. In an SVM, the linear decision function can be thought of as a hyperplane dividing the multidimensional voxel-space into two classes, and voxel weights are learned so as to maximize the distance between the hyperplane and the closest observed example. We conducted binary classification with a linear kernel, using a fixed regularization parameter (C=1) to control the tradeoff between margin size and training error. We restricted ourselves to linearly decodable-signal under the assumption that a linear kernel implements a plausible readout mechanism for downstream neurons (Seung & Sompolinsky, 1993; Hung et al., 2005; Shamir & Sompolinsky, 2006). Given that the brain likely implements nonlinear transformations, linear-separability within a population can be thought of as a conservative but reasonable estimate of the information available for explicit readout (Dicarlo & Cox, 2007).

For each classification, the data were partitioned into multiple cross-validation folds where the classifier was trained iteratively on all folds but one, and tested on the remaining fold. Classification accuracy was then
averaged across folds to yield a single classification accuracy for each subject in the ROI. A one-sample t-test was then performed over these individual accuracies, comparing to chance classification of 0.50 (all t-tests on classification accuracies were one-tailed). While parametric tests are not always appropriate for assessing the significance of classification accuracies (Stelzer et al., 2012) the assumptions of these tests are met in the present case: the accuracy values are independent samples from separate subjects (rather than individual folds trained on overlapping data), and the classification accuracies were found to be normally distributed around the mean accuracy. For within-stimulus analyses (classifying within facial expressions and within situation stimuli), cross-validation was performed across runs (i.e. iteratively train on 7 runs, test on the remaining 8th). For cross-stimulus analyses, the folds for cross-validation were based on stimulus type. To ensure complete independence between training and test data, folds for the cross-stimulus analysis were also divided based on even vs. odd runs (e.g. train on even run facial expressions, test on odd run situations).

**Wholebrain searchlight classification**: The searchlight procedure was identical to the ROI-based procedure except that the classifier was applied to voxels within searchlight spheres rather than individually localized ROIs. For each voxel in a gray matter mask, we defined a sphere containing all voxels within a 3-voxel radius of the center voxel. The searchlight size (123 voxels) was selected to approximately match the size of the regions in which effects were found with the ROI analysis, and we again conducted an ANOVA to select the 80 most active voxels in the sphere. Classification was then performed on each cross-validation fold, and the average classification accuracy for each sphere was assigned to its central voxel, yielding a single accuracy image for each subject for a given discrimination. We then conducted a one-sample t-test over subjects’ accuracy maps, comparing accuracy in each voxel to chance (0.5). This yielded a group t-map, which was assessed at a p<.05, FWE corrected (based on SPM’s implementation of Gaussian Random Fields).

**Whole brain random effects analysis (univariate)**: We also conducted a whole-brain random effects analysis to identify voxels in which the univariate response differentiated positive and negative valence for faces and for situations. The conjunction of these two contrasts would identify voxels in which the magnitude of response was related to the valence for both stimulus types.
1.3 RESULTS

1.3.1 Experiment 1

1.3.1.a Regions of interest

Using the contrast of Belief > Photo, we identified 7 ROIs in each of the 21 subjects: rTPJ, ITPJ, rATL, PC, DMPFC, MMPFC, VMPFC, and using the contrast of faces > objects we identified right lateralized face regions OFA, FFA, and mSTS in 18 subjects (out of 19 subjects who completed this localizer).

1.3.1.b Multivariate results

*Multimodal regions (pSTC and MMPFC)*: For classification of emotional valence for facial expressions, we replicated the results of Peelen and colleagues (2010) with above chance classification in MMPFC (M(SEM)=0.534(0.013), t(18)=2.65, p=0.008; Figure 1.4) and lpSTC (M(SEM)=0.525(0.010), t(20)=2.61, p=0.008; Figure 1.5). Classification in rpSTC did not reach significance at a corrected (0.05/3) threshold (M(SEM)=0.516(0.007), t(20)=2.23, p=0.019). Note that while the magnitude of these effects is small, these results reflect classification of single event trials, which are strongly influenced by measurement noise. Small but significant classification accuracies are common for single trial within-category distinctions (Eger et al., 2008; Harry et al., 2013, Anzellotti et al., 2013).
The key question for the present research is whether these regions contain neural codes specific to overt expressions or whether they also represent the valence of inferred emotional states. When classifying valence for situation stimuli, we again found above chance classification accuracy in MMPFC ($M(SEM)=0.553(0.012)$, $t(18)=4.31, p>0.001$). We then tested for stimulus-independent representations by training on one kind of stimulus and testing on the other. Consistent with the existence of an abstract valence code, MMPFC supported above chance valence classification across both stimulus types ($M(SEM)=0.524(0.007)$, $t(18)=3.77, p=0.001$). In contrast, lpSTC did not perform above chance when classifying the valence of situation stimuli ($M(SEM)=0.512(0.011)$, $t(20)=1.06, p=0.152$), nor when requiring generalization across stimulus type ($M(SEM)=0.500(0.008)$, $t(20)=0.04, p=0.486$). To directly compare accuracy in lpSTC when classifying within facial expression stimuli and when generalizing across stimulus types, we conducted a paired sample t-test (one-
tailed) comparing classification accuracy for faces to accuracy for cross-stimulus classification: classification accuracy was significantly higher for faces compared to cross-stimulus classification (M=0.525, M=0.500, t(20)=2.00, p=0.029).

Theory of Mind Regions: We performed these same analyses in 6 remaining theory of mind regions (at a corrected alpha of 0.05/6, 0.008). In DMPFC (Figure 1.4), we observed results very comparable to those observed in MMPFC: above chance classification of facial emotion (M(SEM)=0.539(0.016), t(18)=2.39, p=0.014), of emotion from situations (M(SEM)=0.570(0.013), t(18)=5.38, p<0.001), and when generalizing across stimulus types (M(SEM)=0.532(0.008), t(18)=3.95, p<0.001). VMPFC did not perform above chance at a corrected threshold (p<.008) when classifying facial expressions (M(SEM)=0.525(0.009), t(17)=2.62, p=0.009) or situation stimuli (M(SEM)=0.524(0.012), t(17)=1.98, p=0.032); however, cross-stimulus decoding was above chance (M(SEM)=0.527(0.007), t(17)=3.79, p=0.001).

None of the other theory of mind regions classified above threshold when distinguishing positive and negative facial expressions (rTPJ: M(SEM)=0.501(0.010), t(20)=0.06, p=0.478, rTPJ: M(SEM)=0.521(0.012), t(20)=1.85, p=0.040, rATL: M(SEM)=0.525(0.012), t(20)=2.05, p=0.027, PC: M(SEM)=0.514(0.011), t(20)=1.32, p=0.102), positive and negative situations (rTPJ: M(SEM)=0.528(0.014), t(20)=2.04, p=0.027, ITPJ: M(SEM)=0.515(0.009), t(20)=1.57, p=0.066, rATL: M(SEM)=0.510(0.012), t(20)=0.80, p=0.216, PC: M(SEM)=0.523(0.012), t(20)=1.84, p=0.040), or when generalizing across stimulus types (rTPJ: M(SEM)=0.503(0.007), t(20)=0.45, p=0.330, ITPJ: M(SEM)=0.509(0.007), t(20)=1.38, p=0.092, rATL: M(SEM)=0.510(0.006), t(20)=1.85, p=0.039, PC: M(SEM)=0.495(0.008), t(20)=0.60, p=0.724).

Face-selective cortex (Figure 1.5): For valence in facial expressions, we also performed a secondary analysis in face-selective regions rOFA, rFFA, and rmSTS (at a corrected threshold of 0.05/3). We replicated previous reports (Said et al., 2010a,b, Furl et al., 2012, Harry et al., 2013) with classification accuracies significantly above chance in rmSTS (M(SEM)=0.539(0.007), t(14)=5.20, p<0.001) and in rFFA (M(SEM)=0.531(0.012), t(14)=2.59, p=0.011); classification in rOFA did not survive correction for multiple comparisons (M(SEM)=0.529(0.016), t(13)=1.87, p=0.042). For the situation stimuli, the rFFA failed to classify
valence when it was inferred from context (rFFA: M(SEM)=0.508(0.016), t(14)=0.54, p=0.300). In the rmSTS, on the other hand, there was reliable information about situation stimuli in addition to the face stimuli (M(SEM)=0.537(0.014), t(14)=2.57, p=0.011). However, neither region supported above chance cross-stimulus classification (rFFA: M(SEM)=0.499(0.006), t(14)=-0.16, p=0.563; rmSTS: M(SEM)=0.499(0.008), t(14)=-0.17, p=0.565) and classification accuracy was reliably higher (one-tailed test) when training and testing on faces compared to when requiring generalization across stimulus types in rmSTS (M=0.539, M=0.499, t(14)=4.52, p<0.001) and in rFFA (M=0.531, M=0.499, t(14)=2.26, p=0.020).

**Figure 1.5.** Classification accuracy for facial expressions (green), situation stimuli (blue), and when training and testing across stimulus types (red). Cross-stimulus accuracies are the average of accuracies for train facial expression/test situation and train situation/test facial expression. Chance=0.50.

**Follow-up Analyses:**

Given successful valence decoding in dorsal and middle MPFC, we conducted several follow-up analyses to examine the scope and generality of these effects. For facial expressions, we performed cross-validation across the orthogonal dimension of face gender. Both regions of MPFC performed above chance (DMPFC: ...)
indicating that the valence-specific voxel patterns generalize across two face sets that differed at the level of exemplars, identity, and gender. We also tested for generalization across face sets in the remaining regions that supported decoding of facial expressions (rmSTS, rFFA, lpSTC). The neural patterns generalized across the male and female face sets in the rmSTS (M(SEM)=0.524(0.012), t(14)=2.02, p=0.032) but not in rFFA (M(SEM)=0.512(0.012), t(14)=1.00, p=0.167) or lpSTC (M(SEM)=0.509(0.009), t(20)=1.05, p=0.154).

For situation stimuli, both regions of MPFC were able to classify valence across the orthogonal dimension—social vs. nonsocial situations (DMPFC: M(SEM)=0.552(0.012), t(18)=4.44, p<0.001; MMPFC: M(SEM)=0.543(0.011), t(18)=3.97, p<0.001). Finally, to test for possible asymmetry in the cross-stimulus classification, we separated the cross-stimulus analysis into training on faces/testing on situations and training on situations/testing on faces. We observed above chance classification for both train/test partitions in both DMPFC (testing on faces: M(SEM)=0.523(0.011), t(18)=2.18, p=0.021; testing on situations: M(SEM)=0.540(0.007), t(18)=5.47, p<0.001) and MMPFC (testing on faces: M(SEM)=0.525(0.006), t(18)=4.13, p<0.001; testing on situations: M(SEM)=0.524(0.009), t(18)=2.64, p=0.008).

In sum, it appears that dorsal and middle subregions of MPFC contain reliable information about the emotional valence of a stimulus when the emotion must be inferred from the situation, and that the neural code in this region is highly abstract, generalizing across diverse cues from which an emotion can be identified. In contrast, although both rFFA and the region of superior temporal cortex identified by Peelen et al (2010) contain information about the valence of facial expressions, the neural codes in those regions do not appear to generalize to valence representations formed on the basis of contextual information. Interestingly, the rmSTS appears to contain information about valence in faces and situations, but does not form a common code that integrates across stimulus type.

1.3.1.c Whole brain analyses

To test for any remaining regions that may contain information about the emotional valence of these stimuli, we conducted a searchlight procedure, revealing striking consistency with the ROI analysis (see
Supplemental Table 1; Figure 1.6). Only DMPFC and MMPFC exhibited above chance classification for faces and contexts, and when generalizing across these two stimulus types. In addition, for classification of facial expressions alone, we observed clusters in occipital cortex. Clusters in the other ROIs emerged at a more liberal threshold (rOFA, rmSTS at p<.001 uncorrected, rFFA, rpSTC, lpSTC at p<.01). By contrast, whole brain analyses of the univariate response revealed no regions in which the mean response distinguished between positive and negative facial expressions, or between positive and negative contexts (at p<.05, FWE correction based on Gaussian random fields).

![Figure 1.6](image)

**Figure 1.6.** Whole brain, Experiment 1: Classification in whole brain searchlight (sphere radius=3 voxels). p<.05 (FWE corrected using Gaussian random fields)

### 1.3.2 Experiment 2

The results of Experiment 1 suggest that DMPFC and MMPFC contain abstract, stimulus-independent information about emotional valence of perceived and inferred emotions. How is this region related to the regions of MPFC typically implicated in processing value and/or subjective experience? For Experiment 2, we first used a group anatomical mask (Clithero & Rangel, 2013; Bartra et al., 2013) to identify a region of OFC/VMPFC previously implicated in reward/value processing. Consistent with previous reports (Kable &
Glimcher, 2007; Chib et al., 2009), this region showed an overall magnitude effect for positive > negative rewards ($t(15)=3.20, p=0.006$, see Figure 1.7), and could classify positive vs. negative reward trials reliably above chance ($M(SEM)=0.542(0.020), t(15)=2.09, p=0.027$). Interestingly, this canonical reward region did not reliably distinguish positive and negative situations for others ($M(SEM)=0.521(0.018), t(15)=1.15, p=0.135$), and there was no evidence for a common valence code generalizing across self and other ($M(SEM)=0.512(0.014), t(15)=0.80, p=0.219$). Classification accuracies were significantly higher when discriminating self reward values compared to when generalizing across reward and situation trials ($M=0.542, M=0.512, t(15)=1.90, p=0.038$, one-tailed).

![Figure 1.7. OFC/VMPFC. Results from anatomical OFC/VMPFC reward ROI (Clithero and Rangel, 2013; Bartra et al., 2013). On the left: classification accuracy for reward outcomes (purple), situation stimuli (blue), and when training and testing across stimulus types (red). Chance=0.50. On the right: mean beta values in the ROI for each stimulus condition.](image)

What about the regions implicated in abstract valence representation in Experiment 1? By decoding valence within the situation stimuli, we replicate the finding of Experiment 1 that DMPFC and MMPFC contain information about the emotion attributed to a target even when that emotion must be inferred from context (DMPFC: $M(SEM)=0.543(0.021), t(15)=2.04, p=0.030$, MMPFC: $M(SEM)=0.536(0.019), t(15)=1.95, p=0.035$, see Figure 1.8). Do we observe these same neural patterns on trials in which subjects evaluate their own
subjectively experienced emotions? In MMPFC, we observed above chance valence classification for reward trials (\(M(\text{SEM})=0.539(0.018), t(15)=2.17, p=0.023\)) in addition to situation trials. Moreover, neural patterns generalized across positive/negative situations and positive/negative outcomes for the self (\(M(\text{SEM})=0.526(0.010), t(15)=2.60, p=0.010\)). In dorsal MPFC, in contrast, we observed similar classification of the valence of reward outcomes (\(M(\text{SEM})=0.544(0.025), t(15)=1.74, p=0.051\)), but this region failed to classify above chance when generalizing across self and other (\(M(\text{SEM})=0.514(0.013), t(15)=1.07, p=0.150\)).

![Figure 1.8. MPFC, Experiment 2: Classification accuracy for reward outcomes (purple), situation stimuli (blue), and when training and testing across stimulus types (red). Cross-stimulus accuracies are the average of accuracies for train reward/test situation and train situation/test reward. Chance=0.50.](image)

1.4 DISCUSSION

Are there neural representations of emotions that generalize across diverse sources of evidence, including overt emotional expressions and emotions inferred from context alone? In present study, we identified regions in which voxel-wise response patterns contained information about the emotional valence of facial expressions, and a smaller number of regions that distinguished the valence of emotion-eliciting situations. Our results, together with existing literature (Peelen et al., 2010), provide candidate neural substrates for three levels of representation: *modality-specific representations* bound to perceptual invariants in the input, intermediate *multimodal representations* that generalize across canonical perceptual schemas, and *conceptual representations* that are fully
invariant to the information used to identify emotions.

1.4. A Conceptual representations

In DMPFC/MMPFC, we decoded emotional valence from facial expressions, and from animations depicting emotion-eliciting situations. Like other domains of high-level cognition, emotion knowledge is theory-like (Gopnik & Wellman, 1992; Carey, 1985), requiring abstract concepts (e.g. of goals, expectations) to be integrated in a coherent, causal manner. The present results suggest that valence representations in DMPFC/MMPFC are elicited by such inferential processes. We could classify valence when training on faces and testing on situations (and vice versa), replicating the finding that emotion representations in MMPFC generalize across perceptually dissimilar stimuli (Peelen et al., 2010). Moreover, our results demonstrate an even stronger form of generalization: perceived emotions and emotions inferred through generative, theory-like processes activate similar neural patterns in DMPFC/MMPFC, indicating a mechanism beyond mere association of co-occurring perceptual schemas. Thus, the MPFC may contain a common neural code that integrates diverse perceptual and inferential processes to form abstract representations of emotions.

Previous research leaves open the question of whether activity in MPFC reflects mechanisms specific to emotion attribution or mechanisms involved in value or valence processing more generally. In Experiment 2, we found evidence for both kinds of representations. First, we found that the region of OFC/VMPFC implicated in reward processing (anatomical ROI from Bartra et al., 2013; Clithero & Rangel, 2013) does not contain information about the valence of attributed emotions. Second, we found no evidence for a shared representation of experienced and attributed emotion in dorsal MPFC. Finally, in MMPFC, we observed neural patterns that generalized across attributed and experienced emotional events. One interpretation of this result is that attributing positive or rewarding experiences to others depends on general-purpose reward representations that code value in social and nonsocial contexts (Chib et al., 2009; Lin et al., 2012, Ruff & Fehr, 2014). Alternatively, neural responses in MMPFC could reflect the participant’s own empathic reaction to the depicted experiences (e.g. witnessing someone achieve a goal elicits positive emotions in participants). If so, the participant’s empathic reaction might be causally involved in the process of attributing emotions to others (consistent with “simulation
theory”, Goldman and Sripada, 2005; Niedenthal, 2007) or might be a downstream consequence of attribution. Previous results do indicate a causal role for MPFC in emotion perception and attribution: damage to MPFC is associated with deficits in emotion recognition (Shamay-Tsoory et al, 2003; 2009), and direct disruption of MPFC via TMS has been shown to impair recognition of facial expressions (Harmer et al., 2001; see also Mattavelli et al., 2011). Moreover, the degree to which MPFC is recruited during an emotion attribution task predicts individual differences in the accuracy of emotion judgments (Zaki et al., 2009). Future research should continue to distinguish the specific contents of attributed emotions from the emotional response of the participant. For example, can patterns in MPFC be used to classify the attribution of more specific emotions that are unlikely to be shared by the observer (e.g. loneliness versus regret)?

1.4.b Modality-specific representations

In face-selective regions (rFFA, and rmSTS), we found that neural patterns could distinguish positive and negative facial expressions, replicating previous reports of emotion-specific neural representations in these regions (Fox et al., 2009; Xu & Biederman, 2010; Said et al., 2010a,b; Furl et al., 2012; Harry et al., 2013). Neural populations could distinguish facial expressions by responding to relatively low-level parameters that differ across expressions, by extracting mid-level invariants (e.g. eye motion, mouth configuration) that generalize across within-modality transformations (e.g. lighting, position), or by computing explicit representations of facial emotion that integrate multiple facial parameters. The present study used naturalistic stimuli that varied in lighting conditions, face direction, and face position, and found reliable generalization across male and female face sets in rmSTS. Thus, it is possible that these neural patterns distinguish facial expressions based on representations invariant to certain low-level transformations (see Anzellotti et al., 2013). Future research should investigate this possibility by systematically testing the generalization properties of neural responses to emotional expressions across variation in low-level dimensions (e.g. face direction) and higher-level dimensions (e.g. generalization from sad eyes to a sad mouth). Interestingly, the rmSTS also contained information about emotional valence in situation stimuli, but the neural patterns did not generalize across these distinct sources of evidence, suggesting two independent valence codes in this region.
1.4.c Multimodal representations

We also replicate the finding that pSTC contains information about the emotional valence of facial expressions (Peelen et al., 2010). However, unlike DMPFC/MMPFC, we find no evidence for representations of emotions inferred from situations. Interestingly, Peelen and colleagues found that the pSTC could decode emotional expressions across modalities (faces, bodies, voices), suggesting that this region may support an intermediate representation that is neither fully conceptual nor tied to specific perceptual parameters. For example, pSTC could be involved in pooling over associated perceptual schemas, leading to representations that generalize across diverse sensory inputs but do not extend to more abstract, inference-based representations. This interpretation would be consistent with the region’s proposed role in cross-modal integration (Stevenson & James, 2009; Kreifelts et al., 2009). Thus, the present findings reveal a novel functional division within the set of regions (pSTC and MMPFC) previously implicated in multi-modal emotion representation (Peelen et al., 2010).

1.4.d Open questions

While these data provide important constraints on the levels of representation associated with different regions, important questions remain open. First, do the regions identified here contain information about more fine-grained emotional distinctions beyond valence? Previous studies have successfully decoded a larger space of perceived emotions in MMPFC, STS, and FFA (Peelen et al., 2010, Said et al., 2010a,b; Harry et al., 2013). For emotions inferred from context, the neural representation of more fine-grained emotional distinctions (e.g. inferring sadness vs. fear) will be a key question for future research.

This study also leaves open the role of other regions (e.g. amygdala, insula, inferior frontal gyrus) that have previously been associated with emotion perception and experience (Shamay-Tsoory et al., 2009, Singer et al., 2009, Pessoa & Adolphs, 2010). What is the precise content of emotion representations in these regions, and do they contribute to identifying specific emotional states in others? With the searchlight procedure, we found little evidence for representations of emotional valence outside the a priori ROIs. However, whole brain analyses are less sensitive than ROI analyses, and while multivariate analyses alleviate some of the spatial constraints of univariate methods, they still tend to rely on relatively low-frequency information (Op de Beeck, 2010; Freeman
et al., 2011), meaning that MVPA provides a lower-bound on the information available in a given region (Kriegeskorte & Kievet, 2013). Neurophysiological studies (Gothard et al., 2007; Hadj-Bouziane et al., 2012) may help to elucidate the full set of regions contributing to emotion attribution.

Relatedly, how does information in these different regions interact during the process of attribution? A tempting speculation is that the regions described here make up a hierarchy of information flow (e.g. modality-specific face-selective cortex ⇔ multi-modal pSTC ⇔ conceptual MPFC; Adolphs 2002; Ethofer et al., 2006). However, additional connectivity or causal information (e.g. Friston et al., 2003; Bestmann et al., 2008) would be required to confirm such an account, and to directly map different representational content onto discrete stages.

Finally, these findings are complementary to previous investigations of semantic representations (e.g. object categories: Fairhall & Caramazza, 2013; Devereux et al., 2013), which have identified modality-specific representations (e.g. in visual cortex) and representations that generalize across modalities (e.g. across words and pictures in left middle temporal gyrus). The present findings highlight a distinction between representations that are multimodal and those that are based on theory-like causal inferences. Does this distinction apply to other domains, and can it help to clarify the neural organization of abstract knowledge more broadly?

1.4.e General conclusions

The challenge of emotion recognition demands neural processes for exploiting different sources of evidence for others’ emotions, as well as a common code for integrating this information to support emotion-based inference. Here, we demonstrate successful decoding of valence for emotional states that must be inferred from context as well as emotions directly perceived from overt expressions. By testing the scope and generality of the responses in different regions, we provide important constraints on possible computational roles of these regions, and begin to elucidate the series of representations that make up the processing stream for emotional perception, attribution, and empathy. Thus, the present research provides a step towards understanding how the brain transforms stimulus-bound inputs into abstract representations of emotions.
2 Preverbal infants identify emotional reactions that are incongruent with goal outcomes

2.0 ABSTRACT

Identifying the goal of another agent’s action allows an observer to make inferences not only about the outcomes the agent will pursue in the future and the means to be deployed in a given context, but also about the emotional consequences of goal-related outcomes. While numerous studies have characterized the former abilities in infancy, expectations about emotions have gone relatively unexplored. Using a violation of expectation paradigm, we present infants with an agent who attains or fails to attain a demonstrated goal, and reacts with positive or negative affect. Across several studies, we find that infants’ attention to a given emotional display differs depending on whether that reaction is congruent with the preceding goal outcome. Specifically, infants look longer at a negative emotional display when it follows a completed goal compared to when it follows a failed goal. The present results suggest that infants’ goal representations support expectations not only about future actions but also about emotional reactions, and that infants in the first year of life can relate different emotional reactions to conditions that elicit them.
2.1 INTRODUCTION

Success in a social environment depends on capacities to understand, anticipate, coordinate with, and learn from the behavior of others. Human adults readily solve these problems by relying on intuitive knowledge of other minds that specifies the causal relationships linking various mental states to each other, to events or conditions in the external environment, and to overt action (Wellman, 1990; Gelman & Wellman, 1992; Gopnik & Meltzoff, 1997; Carey, 1985). On this basis of this knowledge, a perceiver can recover goals and other mental states from observed behavior (Baker et al., 2008; 2011), and recruit these mental state representations for a range of inferences. For example, we rely on goals or intentions to socially evaluate other agents (e.g. Cushman et al., 2006, Young & Saxe, 2009), to interpret speech and other communicative acts (e.g. Goodman & Stuhlmüller, 2012; Smith, Goodman & Frank, 2013), and to reason about an agent’s emotional state in different contexts (e.g. Parkinson, 2007; Siemer & Reisenzein, 2007; Zaki, Bolger & Ochsner, 2009).

The present research probes the development of this last set of inferences, specifically the ability to predict the emotional consequences of goal-related outcomes. Prior studies using verbal vignettes and pictorial scenarios suggest that young children can identify how a target will feel in response to a particular event (Wellman and Wooley, 1990): by 2-3 years of age, children reason about emotions as well as desires and preferences, inferring others’ emotional states in the absence of overt reactions (e.g. Wellman & Banerjee, 1991; Wellman & Bartsch, 1988; Russell, 1990; Yuill, 1984; for related findings with younger children, see Vaish, Carpenter & Tomasello, 2009; Chiarella & Poulin-Dubois, 2013). To investigate the origins of this knowledge, the present research examines whether basic emotion attribution abilities might be evident in preverbal infants.

By midway through first year of life, humans attend to the intentional movements of others and appear to encode goal-relevant properties of these movements, such as the objects to which they are directed, over more superficial properties, such as their trajectories (Gergely et al., 1995; Woodward, 1998). On the basis of observed actions, infants form expectations both about the outcome of future actions (Woodward, 1998; Jovanovic et al., 2007; Bíró & Leslie, 2007; Csibra et al., 2003) and about the means that will be exploited under different physical constraints (Gergely et al., 1995; Kamewari et al., 2005; Phillips & Wellman, 2005). One interpretation of these
and other findings (Luo & Baillargeon, 2005; Luo & Johnson, 2009; Kovács et al., 2010) is that infants exploit abstract principles to make sense of the movements of others, integrating several relevant variables (outcomes, paths, physical obstacles and barriers to perception) to identify an agent’s goal and anticipate future behavior. On this view, early representations of goal-directed behavior are embedded in a coherent inferential framework for predicting and explaining action (Luo & Baillargeon, 2010; Bíró, Verschoor & Coenen, 2011; Carey, 2009).

Others have avoided appeal to abstract inferential principles, explaining these phenomena in terms of domain-general associative or statistical learning mechanisms operating over sensory or motoric representations (e.g. Paulus, 2012; Paulus et al., 2011; Rakison, Cicchino & Hahn, 2007). In fact, some have argued that infants could exhibit expectations about the path of an action in these experiments without having any representation of the action as goal-directed (Paulus et al., 2011). Moreover, even among theories that grant abstract goal knowledge to infants, early accounts posited a relatively limited inferential mechanism; Gergely and Csibra, for example, proposed that infants represent actions by assuming a teleological stance, analyzing the path an entity takes, the outcomes it achieves, and the physical constraints of the environment in accord with an assumption that actions are efficient with respect to goals (Gergely et al., 1995; Gergely & Csibra, 2003). This mechanism, at least as initially described, would operate over observable variables to form an abstract action representation, but would not posit subjective epistemic states, or other internal psychological states such as emotions.

One way to distinguish between these possibilities is to examine the range of inferences supported by early goal-representations. Upon observing a goal-directed action, are infants’ predictions limited to the path a subsequent action will take and the end state it will achieve, or do infants form a broader set of expectations? In particular, the present research explores whether preverbal infants have expectations about the affective states that are likely to result from different goal outcomes. Despite decades of research on infants’ abilities to process and interpret emotional displays (e.g. Nelson, 1987; Field, et al., 1983; Walker-Andrews, 1997; Moses et al., 2001; Grossman, 2010), there is little evidence to date that infants have knowledge of the eliciting conditions for different emotions. In fact, several findings suggest that young infants might fail to understand the relations between goals and emotions.
First, Repacholi and Gopnik (1998) found that whereas 18-month-old toddlers could use an agent’s positive emotional expression towards a food item to guide their sharing behavior (see also Egyed, Király & Gergely, in press), 14-month-olds ignored the target’s expressed emotion and provided her with the item they themselves preferred. However, this failure could have resulted from conflict between the partner’s preference and the child’s own preference, which must be suppressed in order to help according to the partner’s desire. To eliminate these demands, Vaish and Woodward (2009) used a looking time paradigm investigating whether infants this age could use an agent’s emotional expression to predict her subsequent action. Specifically, infants viewed an agent direct attention and emotion towards one of two objects, and then reach either towards the attended or unattended object. Fourteen-month-old infants looked longer when the agent reached towards the unattended object, regardless of whether her expressed emotion had been positive or negative. The authors interpret this pattern as evidence that these infants did not understand the relation between emotion and goal-directed action. Because emotion cues conflicted with attentional cues, however, it is possible that infants failed to use emotional information because another salient and relevant cue was provided. Infants might nevertheless represent the relations between emotions and goals by this age, and exhibit such understanding in contexts that eliminate these competing demands.

Thus, despite the abundance of research on action understanding in infancy, additional research is needed to characterize the full scope of early goal knowledge, and the trajectory of developmental change in these abilities. In the present studies, we begin to fill this gap by investigating whether preverbal infants form expectations about emotional reactions to goal-relevant outcomes. If infants can represent the affective consequences of achieving or failing to achieve a goal, they should be sensitive to whether an agent’s emotional reaction is consistent with an observed outcome. We were particularly interested in infants between 8 and 10 months of age, given conflicting reports of change and continuity in social cognitive abilities through this range (see Tomasello et al., 2005, Beier and Spelke, 2012; Woodward, 2003, Senju and Csibra, 2008; Luo, 2010).

2.2. EXPERIMENT 1
In Experiment 1, we familiarized infants with events in which an agent pursued a goal (reaching a mat in a particular location) by repeatedly moving to that location, modifying its path based on the constraints of the environment. We then presented events in which the agent either successfully completed or failed to complete this goal, and exhibited an emotional response that was congruent or incongruent with the outcome. If infants link goal outcomes to emotional reactions, they should exhibit heightened attention to events in which there is a mismatch between the outcome and the target’s affective response.

2.2.1 Experiment 1 Method

2.2.1.a Participants

This study was conducted at the Laboratory for Developmental Studies on Harvard University’s campus. Thirty-two 10 month-old infants (15 females) and thirty-two 8 month-old infants (12 females) were tested. An additional nine infants also were tested but were excluded from the data analysis because of fussiness/inattention (n=4), parental interference (n=1), looking time more than 3 standard deviations above the mean (n=1), equipment failure (n=1), or online coding error (n=2). All the infants were healthy, full-term (at least 36 weeks gestation), and living in the greater Boston/Cambridge area.

2.2.1.b Apparatus

The events presented in this experiment consisted of animated sequences of actions and emotional reactions. The events were presented using Keynote software running on a MacBook Pro computer connected to an LCD projector. Infants viewed the events on a large screen (~1.2mx1.7m) while seated on a parent’s lap in a darkened experimental room, at a viewing distance of approximately 1.5 m. A video camera was mounted below the screen and positioned behind a small hole in a black curtain to achieve a clear image of the infant’s face without creating a visual distraction.

2.2.1.c Displays

Computer-animated events were created using Keynote software. The events involved two simple geometric characters: red and purple circles whose schematic faces had small distinguishing features (e.g. different ears and hair) and could assume different expressions. Infants first viewed emotion-familiarization trials (to
ensure that the emotional displays were not entirely novel when presented during test trials) in which the two agents were presented on opposite sides of the screen. One agent exhibited a single positive reaction, in the form of an upturned smile appearing on the face, accompanied by a child-like giggling sound, and a small bouncing movement. The other agent exhibited a single negative emotional reaction, which involved a downturned frown, as well as an infant crying sound (from stimuli reported in Johnson, Dweck, and Chen, 2007) and a slow, side-to-side rocking movement. These reactions occurred sequentially, each lasting for seven seconds. In a second emotion-familiarization trial, each agent exhibited the opposite reaction from that expressed during the first trial.

Test trials consisted of five goal-familiarization events (see Figure 2.1a) in which one of the two agents appeared and engaged in a goal-directed action of moving towards and stopping on a gray square mat. In the first two of these events, the agent moved in a straight path towards the goal. For the following three goal-familiarization events, a barrier appeared, changing in height on each familiarization, and the agent jumped from off screen, adjusting its jump to the height of the barrier. On the third and fourth familiarization events, the agent successfully jumped over the barrier to reach the goal location. On the fifth goal familiarization, the barrier became even higher, and the agent failed to surpass the barrier, instead hitting it and rolling back to the side of entry. These events occurred in rapid succession.

Infants then viewed a goal-outcome event in which the agent engaged in a second attempt towards the goal, and either completed the goal (surmounting the barrier and reaching the goal location) or failed to complete the goal (hitting the barrier and tumbling back down to the starting point) (see Figure 2.1b). The agent then exhibited one of the two emotional reactions described above (Figure 2.1c)\(^5\). The emotional reaction was shown once, and subjects were excluded for inattentiveness if they did not look to the screen during any part of the emotional display.

\(^5\) To validate these stimuli, we showed the same animations to 72 adults on Amazon’s Mechanical Turk and asked them to rate the “strangeness” of the character’s reaction on a scale from 1 (not at all strange) to 7 (extremely strange). Adults judged the incongruent reactions as more strange for both completed and failed goal videos. Mean(SEM): Completed goal/Positive affect=3.61(0.57), Completed goal/Negative affect=5.00(0.53), Failed goal/Positive affect=5.74(0.33), Failed goal/Negative affect=3.06(0.50)
Figure 2.1. Trial structure for Experiment 1. For each of four trials, infants viewed five goal familiarization events (A), followed by a goal-outcome event (completion or failure, B), followed by an emotional reaction event (positive or negative, C).

2.2.1.d Design

Infants viewed a total of four test trials, each involving 5 goal familiarization events followed by a goal-outcome event and an emotional reaction event. Since subjects might have been confused by a single agent who completed its goal and failed to complete the goal on different trials, two different agents were presented, one who succeeded in both test events and one who failed in both events. Thus, one agent successfully completed its goal
and responded with positive emotion on one trial and negative emotion on the other. In the two remaining trials, the other agent failed to complete its goal and responded with either positive or negative emotion. Thus, each subject viewed all four test trial types: completed-positive (congruent), complete-negative (incongruent), failed-negative (congruent), failed-positive (incongruent). Subjects viewed two reactions of a given emotion (following a failed or completed goal), and two reactions of the opposite valance (maintaining the order of failure and completion). Trials were therefore presented either in the order of incongruent-congruent-congruent-incongruent or congruent-incongruent-incongruent-congruent. Between subjects, we counterbalanced whether the first trial involved a failed or completed goal, whether the first emotional response was positive or negative, which agent exhibited which test trial type, and the order and side of the screen on which emotions were introduced during emotion-familiarization trials.

2.2.1. Procedure

Stimuli were presented to infants by an experimenter behind a curtain in the testing room, and live video of the infants’ face was fed to an adjacent coding room. A second experimenter viewed the infant’s face on a television monitor and coded the infant’s attention to the display by pressing a button when the infant was attending to the screen. Prior to presentation of the displays, the second experimenter was calibrated to the relevant gaze locations by the first experimenter calling the subject’s attention to the middle of the screen and to each of the screen’s edges. The coder’s responses were tracked using the Xhab64 software program (Pinto, 1995), which signaled the experimenter in the testing room to progress to the next trial after a pre-established attentional criterion. Both experimenters, including the experimenter presenting the stimuli, were blind to the visual events presented to the infant, and thus to which trials were congruent or incongruent for a given subject. Caregivers were instructed to keep their eyes closed throughout the entirety of the session. Infants’ attention was called to the screen at the beginning of the session by the experimenter saying “Hi, [baby’s name], look at this!”. For the two emotion-familiarization trials, looking time was recorded from the start of the first emotional vocalization in the event, and continued until the infant had disengaged attention from the screen for 2 consecutive seconds or had reached a maximum of 45 seconds of total looking time. Infants then viewed the test
trials, each involving five brief goal familiarizations followed by a goal-outcome event and an emotional reaction event (see Figure 2.1). During reaction events, looking time duration was again recorded from the start of the emotional vocalization and continued until the infant looked away for 2 seconds or reached 45 seconds of total looking time. This entire sequence was repeated for each of four test trials.

2.2.1 Coding and analyses

In order to present events with trial duration contingent on the infant’s attention, online coding was conducted by a researcher in an adjacent room (blind to condition), as described above. Looking times were then coded offline (also blind to condition), and the latter were used for analysis. Another researcher coded 100% of sessions, and these two offline coding measures were highly correlated, r=0.95. To directly test for bias in the coding, we calculated the difference between the main coder and the reliability coder for each trial, and assigned a positive or negative sign to the difference score depending on whether or not it was in the direction of the hypothesis. These values did not significantly differ from zero (M=-0.179, t(255)=-1.293, p= 0.197). We conducted a repeated measures ANOVA with completion (completed goal vs. failed goal) and congruency (incongruent reaction vs. congruent reaction) as within-subject factors and age group (8 vs. 10 months) as a between-subjects factor.

2.2.2 Experiment 1 Results

At both ages, infants looked longer at the incongruent emotional reactions, an effect driven primarily by longer looking to negative affect following a completed goal (Figure 2.2). The ANOVA revealed a main effect of congruency (F(1, 62)=12.451, p=0.001), with infants looking longer at incongruent emotional reactions (M=13.825) than congruent reactions (M=11.713). There was no interaction between congruency and age group (F(1,62)=0.581, p=0.449), and follow up analyses revealed no main effects of any of the counterbalancing factors (familiarization valence order, familiarization start side, test valence order, and test congruence order). In addition

---

6 For the 10-month-old infants, this set of four test trial types was presented a second time, yielding a total of eight test trials per subject. However, it became clear over the course of testing that eight test trials was too demanding on subject’s attention, many of whom did not complete second test set. All reported analyses in Exp 1 are conducted on the first test set only, and all subsequent studies (including the 8-month-old age group of Exp 1, and the conceptual replication in Exp 3) included only one test set per subject.
to the predicted effect of congruency, there was a trend towards a main effect of completion ($F(1,62)=3.8841$, $p=0.053$). To clarify the nature of this effect, we conducted separate t-tests comparing congruent and incongruent reactions for completed and failed goals separately. There was an effect of congruency for the completed goal test events ($t(63)=3.169, p=0.002$) but not for the failed goal test events ($t(63)=1.103, p=0.274$). Thus, the main effect of congruency appears to be driven by longer looking to the negative emotion following a completed goal. Nevertheless, the congruency x completion interaction was not significant ($F(1,62)=2.191, p=0.144$). To confirm that both age groups exhibit sensitivity to the emotional congruency, we conducted a separate repeated measures ANOVA for each age group and found main effects of congruency in the 10-month-old group ($F(1,31)=4.159$, $p=0.050$) and in the 8-month-old group ($F(1,31)=8.524, p=0.006$). There were no differences in infants’ looking time to the emotion-familiarization trials (Mean(SEM): positive-negative familiarization= 19.64(0.10) seconds, negative-positive familiarization 19.65(0.15) seconds).

![Figure 2.2. Experiment 1 Results. Mean looking time to test trials in Experiment 1 for each age group. Error bars indicate SEM.](image)

2.2.3 Experiment 1 Discussion

In Experiment 1, infants’ looking time to the very same emotional display differed depending on whether the reaction was consistent with the preceding action context. In particular, infants looked longer at a negative
emotional display when it followed successful goal completion, suggesting that infants were sensitive to the mismatch between the situation and the emotional response. We observed no difference between the two age groups studied. Based on these results, we suggest that by 8 months of age infants have some knowledge of the conditions that elicit different emotions in others, and can detect when emotional reactions do not fit with the preceding goal context.

If this interpretation is correct, and infants exhibit differential attention to positive and negative displays based on an analysis of the goal outcome, infants should show this effect only if they are able to identify the agent’s goal during the familiarization phase. To test this prediction, we presented infants with a paradigm in which the test events were identical, but a stable goal could not be inferred from the familiarization trials (see similar controls in Gergely et al., 1995; Csibra et al., 1999). By using the same test displays as Experiment 1, this condition helps to control for various low-level differences between the two test events (i.e. faster downward motion in the failed goal case), and for baseline preferences for one of the two emotional reactions or one of the two outcomes.

2.3. EXPERIMENT 2

In Experiment 2, infants viewed outcome and reaction events identical to those in Experiment 1 (an agent sailing over a barrier and landing on the mat, or colliding with the barrier and tumbling to the ground) but were given no evidence during the familiarization events that the character had a stable goal. Instead of viewing familiarization events in which the character engaged in rational, equifinal movement towards a constant goal, infants were familiarized with events in which the agent moved to different locations on each trial via paths that did not match the environmental constraints. If the results of Experiment 1 depend on infants identifying the agent’s goal and outcomes that are consistent or inconsistent with it, they should show no expectations about emotions in this experiment. Alternatively, if this pattern of results was driven by some low-level property of the displays (e.g. the relationship between the agent’s speed of motion during the outcome event and the reaction
event) or by other differences between the failed goal and completed goal trials, the effect should be maintained in this experiment.

2.3.1 Experiment 2 Method

2.3.1.a Participants

Thirty-two 10 month-old infants (15 females) and thirty-two 8 month-old infants (13 females) participated in this study. An additional eight infants were also tested but were excluded from data analysis because of fussiness/inattention (n=4) or online coding error (n=4). All the infants were healthy, full-term (at least 36 weeks gestation) and living in the greater Boston/Cambridge area.

2.3.1.b Apparatus/Procedure

The apparatus and procedure were identical to those reported for Experiment 1.

2.3.1.c Displays

The outcome and reaction events were identical to those of Experiment 1, but the familiarization events differed. The movements were similar to those in the goal-familiarization events in Experiment 1 (straight or arching paths across the screen), but were not efficient with respect to any stable goal. The movements began and ended in arbitrary, varying locations on each event and were not efficient with respect to environmental constraints (e.g. taking an arched path when no obstacle was present; see Figure 2.3). Subjects then saw the agent begin an arched trajectory across the screen, either sailing over the barrier and landing on the mat, or hitting the barrier and tumbling back down, followed by a positive or negative emotional reaction. These reactions events could be construed as congruent or incongruent with respect to the physical outcome (landing on mat or colliding with barrier), but could not be interpreted in terms of a stable goal of the agent.
2.3.1.d Coding and analyses

The coding procedure was identical to Experiment 1. Another researcher coded 27% of sessions, and these two offline coding measures were highly correlated, $r=0.90$. The principal analysis was as in Experiment 1. A further analysis with the additional factor of experiment (1 vs. 2) compared infants’ test trial looking times across the two experiments.

2.3.2 Experiment 2 Results

At both ages and in both action conditions, infants looked equally at the test events with congruent and incongruent emotional outcomes (Figure 2.4). In contrast to Experiment 1, we found no main effect of congruency ($F(1, 62)=0.585, p=0.447$), with infants looking equally to incongruent emotional reactions ($M=11.702$) and congruent reactions ($M=12.233$). There was no interaction between congruency and age group ($F(1,62)=0.914, p=0.343$), and follow up analyses revealed no main effects of any of the counterbalancing factors (familiarization valence order, familiarization start side, test valence order, and test congruence order). As in Experiment 1, we conducted a separate repeated measures ANOVA for each age group and found no effect of congruency in either the 10-month-old infants ($F(1,31)=1.117, p=0.299$) or the 8-month-old infants.
(F(1,31)=0.027, p=0.870). To compare directly the effect of congruency in Experiments 1 and 2, we conducted a repeated measures ANOVA with event valence (completed vs. failed goal in Experiment 1, pass over barrier vs. hit barrier in Experiment 2) and congruency (congruent vs. incongruent reaction) as within subjects factors and experiment (Experiment 1 vs. Experiment 2) as a between subjects factor. This revealed a significant congruency x experiment interaction (F(1,126)=8.314, p=0.005). Congruency x experiment interactions were also observed when separately analyzing 10-month-old infants (F(1,62)=4.195, p=0.045) and 8-month-old infants (F(1,62)=4.116, p=0.046). Infants’ looking times for the two emotion-familiarization trials did not differ (Mean(SEM)): positive-negative familiarization= 19.89(0.14) seconds, negative-positive familiarization 18.52(.13) seconds).

![Figure 2.4](image)

**Figure 2.4.** Experiment 2 Results. Mean looking time to test trials in Experiment 2 for each age group. Error bars indicate SEM.

**2.3.3 Experiment 2 Discussion**

The results of Experiment 2 suggest that differences in infants’ attention to positive and negative affect following the completed goal events depend on prior identification of the agent’s goal during the familiarization. The previous results are therefore unlikely to have been driven by superficial variables that differed across these test conditions, such as differences in the speeds and directions of the agents’ motions. However, the congruency
effect in Experiment 1 was driven primarily by an effect in the completed goal trials. Experiment 3 was conducted to replicate the results of Experiment 1 and to investigate this possible difference between the failed and completed goal contexts.

### 2.4. Experiment 3

 Experiment 3 investigated whether infants would distinguish congruent from incongruent reactions in a scenario that involved a superficially different goal than Experiment 1 (an object-based rather than a location-based goal) and a different set of environmental constraints. This experiment provided a conceptual replication of Experiment 1, as well as a means of exploring the potential asymmetry between completed and failed goal conditions. In Experiment 1, infants exhibited violation of expectation to the negative emotion following a completed goal, but no response to the positive emotion following a failed goal. Could particular aspects of the goal context used in Experiment 1 explain this pattern? Infants in Experiment 1 viewed a goal familiarization in which the agent failed to achieve the goal but did not react emotionally, and then a test trial in which the agent made a second attempt followed by an emotional response. It is possible that the absence of an emotional response following the failed goal familiarization weakened infants’ expectations about the failed outcome event, either by presenting evidence that the agent’s investment in the goal was weak, or by suggesting that the agent would repeat the action until success. To test this possibility, we removed the failed goal-familiarization event in the present experiment.

The failed action in this experiment also differed from that of Experiment 1 in that the barrier that prevented the agent from reaching the goal appeared after the agent began to move toward the object. Thus, when the agent initiated its path towards the goal, there was no evidence of a physical obstacle. In Experiment 1, the obstacle was visible to the agent throughout the event, such that the agent may have had low expectations about the possibility of obtaining the goal. Given that losses are experienced as more negative when a reward is expected (Schultz, Dayan, & Montague, 1997), we aimed to set up a context in which the agent clearly expected to obtain the goal but was thwarted unexpectedly. In Experiment 3, the agent began moving towards the goal object with no apparent obstacle, and the agent’s action was impeded mid-pursuit by the sudden introduction of a barrier. In all
the outcome events, a large obstacle dropped in front of the agent as it moved towards the goal object. Completed and failed outcomes differed in the location of the object with respect to the obstacle. In failed goal trials, the obstacle fell between the agent and the goal object; in the completed goal trials, the object stood between the agent and the fallen obstacle, and therefore remained accessible to the agent.

2.4.1 Experiment 3 Method

2.4.1.a Participants

Twenty-four 10 month-old infants (15 females) and twenty-four 8 month-old infants (11 females) participated in this study. A larger sample size was used in Experiments 1 and 2 because these experiments were the first investigation in this domain. Given that Experiment 3 was a conceptual replication of the robust effect in Experiment 1, we decided on a smaller sample size (one comparable to other studies using similar methods). An additional nine infants were also tested but were excluded from the data analysis because of fussiness/inattention (n=5), parental interference (n=1), experimenter error (n=2), or online coding error (n=1). All the infants were healthy, full-term (at least 36 weeks gestation) infants living in the greater Boston/Cambridge area.

2.4.1.b Apparatus/Procedure

The apparatus and procedure were identical to those reported for Experiments 1 and 2.

2.4.1.c Displays

The displays of affect during emotional reaction events were identical to those in Experiments 1 and 2. The emotion-familiarizations were similar, but rather than the two agents appearing on either side of the screen, a single agent was presented in the center of the screen during each emotional display. In the goal-familiarization events, an agent engaged in a repeated goal-direct action of moving towards and stopping next to a goal-object (a large ball). There were again four trials, each involving an outcome event and a reaction event, preceded by 5 brief goal-familiarization events (see Figure 2.5a). In the first two goal-familiarizations, an agent moved in a straight path towards the goal. In the following two goal-familiarizations, a barrier appeared and the agent updated its path to move around the barrier, coming to rest next to the goal object. On the fifth familiarization a very large barrier
appeared and the agent successfully jumped over the barrier to reach the goal location. These goal-familiarization events occurred in rapid succession.

During the outcome events (see Figure 2.5b), no barrier was present and the agent initiated a straight path towards the goal object. Then, mid-event, a large obstacle fell from the top of the screen, landing in front of the agent. In both completed and failed events, the agent slowed down and came to rest without contacting the barrier. The only difference between these events was whether the goal object was positioned such that the barrier fell between the agent and the goal-object, preventing the agent from completing its goal, or fell on the far side of the goal object, allowing the agent to complete its goal. The agent then reacted with one of the emotional displays used in Experiments 1 and 2.

Figure 2.5. Trial structure for Experiment 3. For each of four trials, infants viewed five goal familiarization events (A). They then viewed a goal-outcome event (B) in which an obstacle fell in front of the agent (on the failed goal trials, the obstacle blocked access to the object, whereas on the completed goal trials the object was still accessible). The agent then exhibited a positive or negative emotional response.

2.4.1.d Coding and analyses

The coding procedure and analyses were identical to those of Experiments 1 and 2. Another researcher coded 25% of sessions, and these two offline coding measures were highly correlated, r=0.99. We again found that
differences between the main coder and reliability coder were not biased in the direction of the hypothesis (M=0.002, t(47)= 0.022, p=0.983).

2.4.2 Experiment 3 Results

At 10 months, infants’ looking patterns mirrored those of Experiment 1, with longer looking to the incongruent emotional reactions, especially following the successfully completed action (Figure 2.6). At 8 months, in contrast, infants’ looking times did not differentiate between the test events. The ANOVA on looking times revealed no main effect of congruency (F(1, 46)=0.264, p=0.610), and a significant congruency x age group interaction (F(1,46)=6.608, p=0.013). Additional analyses revealed no main effects of any of the counterbalancing factors (familiarization valence order, familiarization start side, test valence order, and test congruence order), and no differences in infants’ looking time for the emotion-familiarization trials (Mean(SEM): positive-negative familiarization= 18.54(0.16) seconds, negative-positive familiarization 18.65(0.19) seconds).

To clarify the nature of the congruency x age group interaction, we conducted a separate repeated measures ANOVA for each age group. There was a main effect of congruency in the 10-month-old infants (F(1,23)=6.446, p=0.018), with longer looking to the incongruent trials (M=14.315) than the congruent trials (M=11.602). As in Experiment 1, this effect was driven by an effect of emotional congruence for the completed goal test events (t(23)=2.211, p = 0.037) but not for the failed goal test events (t(23)=1.148 p= 0.263). However, there was no such effect in the 8-month-old infants (F(1,23)=1.676, p= 0.208). In fact, the means were in the opposite direction with slightly longer looking to the congruent reaction (M=11.554) than the incongruent reaction (M=9.746).
To directly compare the effect of congruency in Experiment 1 to the results of the present experiment, we conducted a separate repeated measures ANOVA for each age group with completion (completed goal vs. failed goal) and congruency (congruent vs. incongruent reaction) as within subjects factors and experiment (Experiment 1 vs. Experiment 3) as a between subjects factor. In 10-month-old infants, this analysis revealed a significant effect of congruency ($F(1,54) = 11.005, p=.002$) and no congruency $\times$ experiment interaction ($F(1,54)= 0.643, p=0.426$). In contrast, there was no main effect of congruency for the 8-month-old infants ($F(1,54)= 0.232, p=0.632$), but a significant congruency $\times$ experiment interaction ($F(1,54)= 7.69, p=0.008$).

### 2.4.3 Experiment 3 Discussion

As in Experiment 1, 10-month-old infants showed heightened attention to an emotional reaction that was incongruent with the preceding action context. Again, infants looked longest to a negative emotional reaction when it followed successful completion of a demonstrated goal. In contrast to Experiment 1, however, we found no sensitivity to the incongruent reaction in the younger age group. One explanation of the differing performance of 8-month-old infants in Experiments 1 and 3 is that younger infants more readily understand the goal context in Experiment 1. There, cues such as physical contact between the agent and the barrier, an actual reversal of the
agent’s trajectory towards the goal, and a failed attempt during the goal familiarization may have made it easier for young infants to understand that the goal persisted even when not attained, and to identify when it had been successfully completed or thwarted. In Experiment 3, the evidence for the goal (or the goal completion) may have been less clear, leading to apparent failure at the younger age. Future research could explore this possibility by directly varying the evidence available for inferring the agent’s goal.

2.5. GENERAL DISCUSSION

Sensitivity to the congruency between an agent’s goal outcome and emotional reaction suggests that 8- and 10-month-old infants relate expressions of affect to their surrounding context. Thus, our findings provide preliminary evidence that preverbal infants are sensitive to the conditions that elicit different emotional reactions, and form expectations about emotional displays based on an analysis of the goals that agents pursue. Furthermore, these experiments raise a number of questions regarding the nature of the representations that support these expectations.

While numerous studies have explored the ability to perceive and learn from others’ emotions (see Grossman, 2010), prior research has left open the possibility that infants understand emotions only as communicative signals conveying objective properties of the world (i.e. which objects are good and bad). The present findings suggest that infants also understand emotional reactions as relating to idiosyncratic preferences or goals of an agent. One possibility is that infants make emotional predictions that are tailored to agent-specific goals because they construe these emotions as subjective internal states (see also Egyed et al., in press). However, it is also possible that infants represent these affective expressions as communicative signals reflecting an agent’s current needs or goals. For example, negative affect could be seen as a general plea for help, which infants could find this surprising in the completed goal trials when no unfulfilled goals are apparent\(^7\). Investigating whether infants represent emotional displays as social-communicative acts, as overt behaviors that occur in particular contexts, or as manifestations of internal states will be an important topic for future research. Either way, it is

\(^7\) Thanks to an anonymous reviewer for pointing out this interpretation.
notable that across Experiments 1 and 2, infants appear to have different expectations about the very same event, depending on what is known about the agent’s goals from its prior actions. When the preceding actions were consistently and efficiently directed towards a goal, infants looked longer if a successful action was accompanied by an incongruent, negative emotion. In contrast, when the preceding actions were not consistently goal-directed or efficient, infants showed no differential attention to the same events. It appears that infants have not simply associated a specific observable event (e.g. surmounting an obstacle) with a specific affective display, but instead rely on prior knowledge about the agent and its goal.

Nevertheless, these studies leave open the possibility that infants’ understanding of actions, goals, and emotions is still developing at 10 months and perhaps undergoes change from 8 to 10 months. For example, there have been conflicting claims as to whether infants this age can infer an agent’s goal when a desired result has not been obtained, and whether they have an understanding of failed goals more broadly. To make sense of goal-directed actions, an observer must be able to represent the discrepancy between an agent’s current state and a goal state, and thus must, in some sense, represent whether or not a goal state has been achieved. However, this ability may be distinct from understanding that an agent can possess a goal that it is unable to fulfill. To test for this knowledge, Brandone and Wellman (2009) presented 8, 10, and 12-month-old infants with a failed goal condition in which a hand reached with an arched trajectory over a barrier to retrieve a ball but fell short of grasping it, rendering the reach unsuccessful. At test, the barrier was removed and the actor either reached directly for the ball and retrieved it, or continued to perform the arched reach, which was no longer efficient with respect to the goal object. By 10 months, but not at 8 months, infants looked longer at the inefficient action, suggesting that they encoded the action as directed towards the object even when the agent had not successfully grasped the object during habituation. Based on these results, Brandone and Wellman argued that 8-month-old infants do not construe intentions as internal states that exist independent of the actions taken to fulfill them, and that a more complete understanding of intentions emerges between 8 and 10 months. An alternative interpretation is that 8-month-old infants do understand that goals can be completed or failed, but simply require more information in order to correctly identify an agent’s goal in a particular instance. The evidence provided by the failed goal
demonstration may have been more ambiguous for infants, such that they identified the goal incorrectly or not at all. Consistent with the latter interpretation, other studies report the ability to infer goals from failed actions in infants younger than 10 months (Hamlin, Newman & Wynn, 2009).

While emotion attribution may be one way to gain traction on the issue of failed goal understanding, the present results do not clearly distinguish between these possibilities. On the one hand, if an understanding of failed goals is still developing towards the end of the first year, this development might explain the fact that infants exhibit expectations about emotions in the completed goal trials but not the failed goal trials. On this interpretation, when the goal is achieved, infants represent this as a positive state and are surprised by a negative emotional response. In the case of a failed attempt, infants may simply represent the goal as ongoing or as having changed, and therefore fail to differentiate between the positive and negative emotional reaction following this event. However, it is also possible that infants comprehend the failed attempts, but do not map them onto the expected emotions, either because they have yet to learn the relationship between failure and negative emotion, or because both emotions are reasonable responses to this event (see discussion below). Future research might test these interpretations by investigating the development of failed goal understanding more broadly, and by exploring expectations about positive and negative emotions in other emotion-eliciting contexts that do not involve thwarted goal-directed actions.

The present studies are also inconclusive with respect to developmental change between 8 and 10 months. Prior research has yielded conflicting reports regarding the continuity of social cognitive abilities in this age range, and the present findings are similarly ambiguous. Although 10-month-old infants made this distinction across two different action contexts, 8-month-old infants succeeded in only one of these contexts. While it is possible that this finding reflects some change in infants’ understanding of emotions, this pattern could also arise from differences in the ease with which infants identify the agent’s goal for the two sets of stimuli. Future research should examine the robustness of infants’ sensitivity to goal-affect relations at these and other ages.

On the basis of Experiments 1-3, we have argued that infants’ attention to the very same emotional display varies based on whether the reaction is congruent with the preceding goal context. However, are there
other interpretations of the reported data? One possibility would be that infants have a baseline attentional preference for the negative emotional display (accounting for the longer looking to the negative compared to positive affect in the completed goal condition), and that this general attentional bias is masked by complexity/confusion in the control trials (Experiment 2) and failed goal trials (Experiment 1 and 3). While we cannot rule out this possibility conclusively, we find it to be a less plausible interpretation of the data for several reasons. First, we included a standard control condition (e.g. Gergely et al., 1995; Csibra et al., 1999) in which test events are identical to those shown in the experimental condition, but where the familiarization phase should not be construed as goal-directed. Given that we find no attentional preference in this condition, one would have to argue that the complexity or novelty of these events overwhelms the expression of such a preference. Specifically, although the movements in Experiment 2 are themselves well-matched to the familiarization events in Experiment 1, these actions are less predictable when not goal-directed. While the unpredictable action events could introduce processing demands that mask a baseline attentional bias for negative affect, this explanation does not readily extend to the failed goal events (in which we again observe no difference between positive and negative affect) as these contain coherent, predictable goal-directed action and are no more complex than the successful goal events. In particular, Experiment 3 is very well matched across the failed and completed goal trials, which differ only in the placement of the goal object with respect to the barrier.

One possibility (as discussed above) is that infants do not understand the failed attempt in these trials, and perhaps construe this event as a goal change. However, if this apparent goal change was sufficiently confusing to overwhelm the possible attentional effect, we should find no effect for any conditions in Experiment 1, as all trials include a failed attempt immediately prior to the outcome. Furthermore, if actions in the failed goal trials were confusing or surprising to infants, we might expect heightened attention to both of these events. Instead, we observe very comparable looking time for the failed goal events and the completion event followed by laughter. It is only the events in which the agent exhibits negative affect following goal completion that elicit heightened attention in this study.
However, even if we reject this leanest possibility, there are a number of open alternatives concerning the scope and depth of infant’s emotion knowledge. Above, we suggested that infants might form a coherent, generative model of an agent’s mind, using behavior to infer an agent’s goal, and representing the emotional states that result from achieving or failing to achieve this goal state. On this view, infants represent several key psychological variables (e.g. goals or preferences, emotional states) and the causal laws that relate them. However, a possibility that remains open in the present research is that infants have more directly associated goal completion and failure with overt expressions of affect (either their own or others’), without inferring any sort of internal emotional state. This account still requires that the infant form a representation of the agents’ goal, and whether it has been attained. However, on this view, the infant does not posit any internal emotional state, but instead maps the abstract outcome representation onto a perceptual emotion schema directly.

The present research does not distinguish between these alternatives. Indeed, this is a challenge faced by all researchers studying theory of mind in nonverbal creatures (see discussion in Penn & Povinelli, 2007; Heyes, 1998; Woodward, 2005; Perner & Ruffman, 2005), and is not one that we can resolve here. With respect to the specific claims of this paper, however, a relatively lean explanation does seem plausible. If, from early in life, infants represent actions in terms of their goals, and can distinguish events in which an agent’s goal state is achieved from events in which it is not, it may be fairly straightforward for infants to learn to associate these outcome variables with the observed facial and vocal expressions of the agents that perform them. Future theoretical and empirical work is needed to distinguish this interpretation from a view in which infants represent a number of causally related internal states in a coherent, theory-like way.

An additional open question concerns the origins of these expectations about emotional reactions. The present results suggest that by 8 months, infants can identify an agent’s goal on the basis of observed behavior and form appropriate expectations about how the agent will then react to completing that goal. However, these results do not bear on the initial origins of these expectations. Infants begin to exhibit sensitivity to others’ action goals as early as 3 months (Sommerville & Woodward, 2005; Luo, 2011, Skerry et al., 2013) and can discriminate different facial expressions starting early in the first year (e.g. Field et al., 1987). One possibility is that as soon as
infants encode the goals of observed actions, they represent the affective consequences of completing these goals. Alternatively, infants might start out with a more restricted schema, similar to that proposed by Gergely and colleagues (1995), and learn over the course of development that failed and completed goals elicit systematically different emotional displays. This learning could take the form described above, where infants map goal outcomes directly onto perceptual representations of emotional displays, or the regularities between outcomes and emotions could support learning over more abstract psychological variables to form theories about the way different mental states interact. The present research cannot distinguish between these possibilities.

Understanding the origins of these expectations might also shed light on the potential asymmetry between failed and completed goals. In the present studies, infants showed violation of expectation to negative affect following a completed goal, but did not distinguish between positive and negative emotion following a failed goal. One explanation, discussed above, is that infants do not have a complete understanding of failed goals. However, this pattern could also be explained in terms of regularities in the input. Humans very rarely exhibit negative affect in response to positive events, but frequently remain neutral, or even laugh, in response to simple failed actions. It seems quite possible, then, that infants receive greater exposure to the correspondence between completed goals and positive emotion than they do the correspondence between failed goals and negative emotions. There is also evidence that beginning in infancy, humans more readily learn from negative information (see Vaish, Grossman, and Woodward, 2008). Thus, it is possible that infants simply learn regularities surrounding negative emotions (that they tend to follow failure, not success) more readily than they do those surrounding positive emotions.

A final outstanding question concerns the relevance of early emotion knowledge to infants’ understanding of, and engagement in, cooperative or prosocial interactions. A number of studies have found that infants preferentially look at, reach towards, and reward ‘helpful’ agents over ‘hindering’ agents: findings that were interpreted as an innate preference for prosocial others (e.g. Kuhlmeier et al., 2003; Hamlin et al., 2007; 2011; Hamlin & Wynn, 2011; but see Scarf et al., 2012). Similarly, as soon as they are physically capable, toddlers themselves engage in actions that complete others’ instrumental goals, and do so with seemingly little regard to
the costs involved or the rewards to be gained (Warneken & Tomasello, 2006; Warneken et al., 2007). A tempting interpretation of these various phenomena is that infants understand the affective value associated with failed and completed goals, and are motivated by the emotional state of the recipient. However, it is unknown whether these preferences and prosocial behaviors are supported by emotion knowledge of the kind investigated here. Given that prosocial behavior is related to empathy and affective perspective-taking in adults (Eisenberg & Fabes, 1990) and young children (Vaish, Carpenter & Tomasello, 2009; Vaish & Warneken, 2012), it would be interesting to test whether these earliest prosocial tendencies also rely upon a developing ability to infer emotions from context.

While the present findings raise many unanswered questions, they nonetheless constitute a first step towards characterizing the nature of infants’ emotion knowledge, and shed light on the scope of their early goal concepts. Interpreting the behavior of other people in terms of underlying goals or intentions is central to learning from and engaging with others. Beginning in infancy, humans appear to represent others’ movements in terms of the goals around which they are structured (Gergely et al., 1995; Woodward, 1998), and these goal representations guide infants’ imitation of others (e.g. Gergely, Bekkering, and Király, 2002; Mahajan & Woodward, 2009), and their social interactions (e.g. Behne, Carpenter, Call & Tomasello, 2005; Warneken & Tomasello, 2006). Here, we provide evidence that infants also form expectations about which affective expressions are likely to follow a successful goal outcome, suggesting that these goal representations may play a relatively flexible role in preverbal infants understanding of others’ behavior.
Neural representations of emotion are organized around abstract event features

3.0 ABSTRACT

Research on emotion attribution has tended to focus on the perception of overt expressions of at most 5 or 6 basic emotions. However, our ability to identify others’ emotional states is not limited to perception of these canonical expressions. Instead, we make fine-grained inferences about what others feel based on the situations they encounter, relying on knowledge of the eliciting conditions for different emotions. In the present research, we provide convergent behavioral and neural evidence concerning the representations underlying these concepts. First, we find that patterns of activity in mentalizing regions contain information about subtle emotional distinctions conveyed through verbal descriptions of eliciting situations. Second, we identify a space of abstract situation features that well captures the emotion discriminations subjects make behaviorally, and show that this feature space outperforms competing models in capturing the similarity space of neural patterns in these regions. Together, the data suggest that our knowledge of others’ emotions is abstract and high-dimensional, that brain regions selective for mental state reasoning support relatively subtle distinctions between emotion concepts, and that the neural representations in these regions are not reducible to more primitive affective dimensions such as valence and arousal.
3.1 INTRODUCTION

The emotional states of others can be identified by a number of cues: we can recognize what someone is feeling based on their facial expressions (Ekman, 1992; Izard, 1971), affective vocalizations (Bachorowski & Owren, 2003; Sauter et al., 2010), or body posture (Aviezer et al., 2012; Dael et al., 2012; de Gelder, 2006). However, we can also attribute subtle emotions based solely on the situation a person encounters (Ortony, 1990), and our vocabulary for attributing these states extends beyond the small set of emotions associated with canonical emotional displays (Fontaine, Scherer, Roesch, & Ellsworth, 2007b; Frijda, 1986; Ortony & Turner, 1990). In many cases, surrounding context has been found to modulate or even dominate the perception of emotion from overt expressions (Barrett, Lindquist, Bliss-Moreau, et al., 2007; Barrett & Kensinger, 2010; Barrett et al., 2011; Hassin et al., 2013).

While the space of emotional states perceived in faces has been studied extensively (Abelson & Sermat, 1962; Ekman & Rosenberg, 1997; Russell & Bullock, 1986b), little is known about how conceptual knowledge of others’ emotions is organized, or how that knowledge is encoded in the human brain. What are the relevant features of events that allow us to make fine-grained emotional attributions (e.g. distinguishing when someone will feel angry versus disappointed, or excited versus proud) and what are the dimensions of the space by which neural populations represent emotions? Here, we argue that emotion attribution recruits a rich theory of the causal context of different emotions, and show that dimensions of this intuitive theory underlie emotion representations in brain regions associated with theory of mind.

3.1.a What neural mechanisms support fine-grained emotion attributions?

As with behavioral research, studies of the neural basis of emotion attribution have generally focused on the perception of affect in facial or vocal expressions (Adolphs, 2002a; Calder & Young, 2005). For example, different facial expressions elicit discriminable patterns of activity in regions of the superior temporal sulcus (STS) and the fusiform gyrus (Furl et al., 2012; Harry, Williams, Davis, & Kim, 2013; Said, Moore, Engell, et al., 2010; Said, Moore, Norman, et al., 2010), while emotional prosody can be decoded in secondary auditory cortex (Ethofer et al., 2009). Some have argued that these overt expressions communicate a set of five or six basic
emotions (Ekman, 1992; Ekman & Rosenberg, 1997; Izard, 1992; Izard, 1994), while other data suggest that faces and voices support even fewer universal discriminations (Aviezer et al., 2008; Jack et al., 2014; Russell, 1994). In addition to regions distinguishing observable displays of emotion, recent research suggests that the medial prefrontal cortex (MPFC) contains representations of emotion that are invariant to perceptual modality (Chikazoe, Lee, Kriegeskorte, & Anderson, 2014; Peelen et al., 2010), generalizing to emotions inferred in the absence of any overt expression (Skerry & Saxe, 2014).

While these studies move beyond modality-specific perceptual representations, they nonetheless focused on relatively coarse distinctions, decoding either valence (Skerry & Saxe, 2014) or 5 basic emotions (Peelen et al., 2010). Does the MPFC also contain representations that support more fine-grained emotional discriminations?

To address this question, we constructed verbal stimuli (see Table 1) describing events or situations that would elicit one of 20 different emotions (validated using 20-AFC behavioral experiment with independent subjects; see Experimental Procedures) and use multi-voxel pattern analysis (Haxby et al., 2001; Kriegeskorte, Goebel, & Bandettini, 2006) to test which regions contain information about these subtle emotional distinctions.

Table 1. Example stimuli

<table>
<thead>
<tr>
<th>Stimulus Type</th>
<th>Example Stimulus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Emotion</td>
<td>After an 18-hour flight, Caitlin arrived at her vacation destination to learn that her baggage (including necessary camping gear for her trip) hadn’t made the flight. After waiting at the airport for 2 nights, Caitlin was informed that the airline had lost her luggage altogether and wouldn’t provide any compensation.</td>
</tr>
<tr>
<td></td>
<td>For months, Naomi had been struggling to keep up with her various projects at work. One week, the company announced that they would be making massive payroll cuts. The next day, Naomi’s boss asked her to come into his office and close the door behind her.</td>
</tr>
<tr>
<td></td>
<td>Linda was having financial difficulties after graduating from college. She worked over-time and lived very meagerly, but still had trouble making her loan payments. One day, she received a letter from her grandfather saying that he wanted to help. A check for $8,000 was enclosed.</td>
</tr>
<tr>
<td></td>
<td>Dana always wanted a puppy, but her parents said it was too much of a hassle. One summer afternoon, Dana’s parents returned from a supposed trip to the grocery store, and Dana heard barking from inside her garage. She opened the door to see her parents holding a golden retriever puppy.</td>
</tr>
<tr>
<td>Physical Pain</td>
<td>One afternoon, Caitlin was running through her house while playing tag with her friend. After going through a doorway, Caitlin slammed the door behind her, but her fingers were caught in the door. When they opened the door, two of her fingers were broken.</td>
</tr>
</tbody>
</table>
As a first step, we train a classifier to distinguish the 20 emotions using distributed patterns of activity across voxels in a region, and test whether the emotion category of a new stimulus can be classified based on the pattern of neural activity it elicits. In addition to whole-brain analyses, we focus on a priori regions of interest (Peelen et al., 2010; Skerry & Saxe, 2014), the strongest candidates being subregions of MPFC—dorsal medial prefrontal cortex (DMPFC) and middle medial prefrontal cortex (MMPFC). However, the MPFC is part of a larger network of regions involved in reasoning about others' mental states (Mitchell, 2009; Saxe & Kanwisher, 2003): the posterior cingulate/precuneus (PC), bilateral temporal parietal junction (TPJ) and bilateral anterior temporal lobes (ATL). These remaining regions of the “Theory of Mind” (ToM) network have been associated with causal attribution of emotion (Bruneau, Pluta, & Saxe, 2012b; Spunt & Lieberman, 2012; Zaki, Weber, Bolger, & Ochsner, 2009), and thus these remaining regions of the “Theory of Mind” (ToM) network serve as additional candidate regions for fine-grained emotion concepts.

We then use Representational Similarity Analysis (RSA: Kriegeskorte et al., 2006b; Kriegeskorte & Kievit, 2013) to characterize emotion representations in ToM brain regions, and test competing hypotheses about the features that best explain that representational space (see Figure 3.1). RSA complements classification analyses by providing a framework for characterizing the representational structure within a region, and for testing competing models of that structure (Kriegeskorte & Kievit, 2013; Kriegeskorte, Mur, & Bandettini, 2008). While above-chance classification of different emotions would demonstrate that a particular region contains information that can differentiates emotions, classification accuracies alone reveal little about the underlying representations. In RSA, neural population codes are represented in terms of the similarity of the neural patterns elicited by different stimuli or conditions. The representational dissimilarity matrix (RDM) of the conditions in a given region can then be compared to the similarity spaces captured by different models (Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte, Mur, & Bandettini, 2008). Importantly, RSA allows for comparison of hypotheses that take different forms (i.e. RDMs can be outputs of formal computational models or derived from quantitative behavioral results) and have different numbers of parameters (the correlation between model and neural RDMs is parameter free, eliminating the risk of over-fitting with more complex models).
RSA Methods: Representational dissimilarity matrices (RDMs) encode the pairwise Euclidean distances between different emotions within each feature space. For each region, a neural RDM captures the pairwise Euclidean distances between different emotions in the patterns of activity elicited across voxels (DMPFC shown here). Feature spaces are fit to the neural data by computing correlations between feature space RDMs and neural RDMs for each region in each subject.
3.1.b Candidate feature spaces for emotion inference

A dominant approach in affective neuroscience has been to represent emotions as points within some low-dimensional space of more basic affective states. One possibility is that the space of emotions is built from a small set of basic emotions (e.g. happiness, sadness, fear, anger, and disgust), each associated with a prototypical facial expression, behavioral profile, and innate neural substrate (Darwin, 1872; Ekman, 1992; Ekman, 1993; Ekman & Cordaro, 2011; Levenson, 2011; Panksepp, 1992; Shariff & Tracy, 2011). On this view, the diverse space of human emotion can be understood as combinations of these more basic emotional states (Oatley & Johnson-Laird, 1996; Plutchik, 2001; Turner, 2000). For example, a recent study found that although human faces could convey as many as 21 discriminable emotional expressions, these emotions could be decomposed into linear combinations of 6 more primitive expressions (Du et al., 2014).

A competing theory of emotional perception is the “circumplex” model, according to which emotions live within a core affective space composed of only two primitive dimensions: valence and arousal (Barrett, 2006b; James A. Russell, 1980; James A. Russell & Bullock, 1986a, 1986b; Watson et al., 1999). Valence and arousal are argued to correspond to two innate systems that are implemented in distinct neural circuits and recruited to varying degrees across different emotions (Barrett & Bliss-Moreau, 2009; Barrett & Wager, 2006; Lindquist, Satpute, et al., 2015; Lindquist et al., 2012; Wilson-Mendenhall et al., 2013). Thus, a second proposal is that neural representations of emotion can be reduced to a linear combination of two these neurophysiological dimensions (Posner et al., 2009).

These theories provide two competing hypotheses about the features or dimensions that structure neural representations of attributed emotion. Although many have focused on the differences between these two proposals (Barrett, 2006a; Posner, Russell, & Peterson, 2005), they have much in common. Both approaches aim to reduce emotions to a smaller number of categories or dimensions, which are assumed to be basic affective states rooted in innate neural substrates. However, much of the empirical support for both proposals comes from studies on the perception of emotions from overt expressions (Ekman, 1992; Engen et al., 1958; Russell & Bullock, 1986a), and from research on the neural correlates of perceiving or inducing different emotions (Barrett & Wager,
While these low-dimensional spaces successfully capture the emotions people perceive in overt expressions, they may be inadequate to account for the full variety of human emotional concepts (Barrett, Lindquist, & Gendron, 2007; Clore & Ortony, 2013; Fontaine et al., 2007b; Ortony, 1990).

Here, we present subjects with the rich causal structure of eliciting situations rather than overt emotional displays. Thus, we hypothesize that the present paradigm will evoke neural representations of emotion that differ both in dimensionality and content from the models that have dominated research on perception of facial expressions. To test this hypothesis, we consider a third feature space that represents emotions in terms of abstract features of the events that give rise to them. According to appraisal theory, emotional reactions relate systematically to people’s interpretations or “appraisals” of the events around them (Ellsworth, 2013; Lazarus, 1991; Scherer, 1999), and there have been various proposals concerning the specific event appraisals that correspond to different emotions (Meuleman & Scherer, 2013; Mortillaro, Meuleman, & Scherer, 2012; Roseman & Smith, 2001; Scherer & Meuleman, 2013). Drawing from this literature, we generated a set of 38 abstract event features thought to reliably vary across different emotions concepts (e.g. Did someone cause this situation intentionally or did it occur by accident? Was the person interacting with other people in this situation? Was this situation a familiar event/situation for the person? See Supplemental Experimental Procedures for full list of appraisal features).

A main goal of the present research is to test whether emotion representations in MPFC and other ToM regions can be well-explained by any of these three candidate feature spaces: the “circumplex” space defined by the judgments of valence and arousal for each stimulus, the “basic emotion” space defined by judgments of the extent to which the stimulus elicited each of 6 basic emotions (happy, sad, angry, afraid, disgusted, or surprised), and the 38-dimensional appraisal space. Importantly, the latter space differs from the other two not only in its

---

8 We drew on appraisal theory to generate event features because researchers in that tradition have been most explicit about what the relevant conceptual dimensions might be. However, our general approach is also compatible with constructivist theories (Brosch, 2013; Clore & Ortony, 2013; Lindquist, 2013; Lindquist, MacCormack, & Shablack, 2015) in which assume that affective primitives like valence and arousal must be combined with abstract conceptual knowledge to differentiate others’ emotions.
dimensionality (38 dimensions vs. 6 or 2) but also in its content: rather than reducing the space of emotions to a smaller set of purportedly "basic" affective states, it aims to encode emotions in terms of abstract features of the causal contexts that tend to elicit them. Of course, the hypothesis that neural representations of emotion concepts are best captured by a high-dimensional space of appraisal features is not at odds with the claim that simpler dimensions like valence and arousal contribute to the organization of our emotion knowledge. For example, the 38-dimensional space contains features such as goal consistency and pleasantness that intuitively relate to the dimension of valence. The question, then, is whether the representations in regions like MPFC can be exhausted by one of the simpler spaces.

To create RDMs that encode each of these feature spaces, an independent group of subjects rated each verbal stimulus on a large set of features (i.e. 38 abstract event features, as well as ratings of 6 basic emotions, and of valence and arousal). These behavioral ratings were used to form three candidate representational spaces (see Figure 3.4), where each emotion category is captured as a feature vector within that space (see Figure S3 for examples of appraisal profiles). First, we examine which space best captures the set of emotions subjects attribute behaviorally by testing whether models trained on each of the relevant feature vectors for a subset of the stimuli could reliably classify the emotion label of untrained stimuli (based on stimulus-specific feature values). Do any of these feature spaces provide a sufficient basis to match performances of human subjects in discriminating these 20 emotions? Then, to test which feature space best explains the neural representation of these stimuli, we compute the similarity of conditions within each proposed feature space and compare the RDMs of candidate models to neural RDMs derived from patterns of activity across voxels in a particular region. Thus, we can test whether regions implicated in emotion discrimination are better characterized by a model containing only dimensions such as valence and arousal (the circumplex space) or 6 basic emotions, or by a representational space defined in terms of abstract event variables.

3.2 RESULTS

3.2.a Classification
In the scanner, subjects read 200 stimuli describing situations that would cause a particular emotion (see Experimental Procedures; example stimuli provided in Table 1). Subjects were instructed to consider how the target would feel in the situation, and rate the intensity of the experience for the target. To confirm that these stimuli elicit reliable fine-grained emotional attributions, a group of subjects on Amazon Mechanical Turk (MTurk: https://www.mturk.com) were asked to choose which of 20 emotion labels best described the emotion of the character in each stimulus. These subjects performed well above chance (relative to the emotion the stimulus was intended to elicit), classifying the stimuli with 65%\(^9\) accuracy (chance=5%; see Figure S2A for behavioral confusion matrix). That is, subjects attribute consistent emotions from these stimuli, providing a benchmark with which to compare different models and brain regions.

To identify regions in which neural patterns contain information about emotions, we first replicate the finding that MPFC contains modality-independent emotion representations by testing whether neural patterns in MFPC can distinguish the valence of these verbal stimuli. We localized MPFC and other regions selective for ToM using individual subject localizers (see supplemental results). We selected a subset of emotion conditions that most closely align with the positive and negative conditions used in previous work (Skerry and Saxe, 2014), and tested whether neural patterns in MPFC would support above-chance classification of these conditions. Replicating prior work, classification of valence was reliably above chance in both DMPFC (M(SEM)=0.610(0.028), t(19)=3.889, p<0.001) and MMPFC (M(SEM)=0.603(0.019), t(19)=5.530, p<0.001).

We then investigate whether these or other regions contain information about the full set of 20 emotions. We conducted a whole brain-searchlight to find regions in which the local neighborhood could classify the 20 emotions above chance. The set of regions that could reliably decode the 20 emotions was largely restricted to regions of the ToM network (particularly DMPFC, RTPJ, LTPJ; see Figure 3.2B and Supplemental Data Table S2). The searchlight analysis exhibits striking overlap with the set of regions recruited for theory of mind (Figure

\(^9\) Note that because the set of emotions included subtle discriminations (apprehensive vs. terrified), and stimulus events could evoke a combination of different emotions, either simultaneously, or over the course of the vignette, we should not expect to obtain 100% agreement across subjects.
3.2B shows overlap between the searchlight (FWE p<.05, k>25) and the random effects analysis of the localizer task, shown at p<.001 uncorrected), and justifies our continued focus on these a priori ROIs.

Consistent with the searchlight results, we were able to classify emotions above chance (1/20 emotions, 5%) based on neural patterns in all individually localized ToM regions: DMPFC: M(SEM)=0.093(0.005), t(19)=9.018, p<0.001; MMPFC: 0.094(0.006), t(19)=7.043, p<0.001; VMPFC: 0.080(0.006), t(17)=5.156, p<0.001; RTPJ: 0.092(0.005), t(21)=8.205, p<0.001; LTPJ: 0.075(0.005), t(21)=4.744, p<0.001;PC: 0.079(0.006), t(21)=4.749, p<0.001;RSTS: 0.082(0.006), t(20)=5.380, p<0.001 (Figure 3.2A).

Figure 3.2. MVPA classification results: A. Above chance 20-way classification of emotions in all ToM regions. B. Whole-brain random-effects analysis of ToM localizer (FB>FP, green); searchlight map for 20-way emotion classification (red); overlap (yellow).

Moreover, in the judgments provided by subjects on MTurk there are reliable differences across the emotion categories in the extent to which subjects make consistent judgments (e.g. that subjects are very reliable in their attribution of “terrified” but less consistent in labeling an event as “joyful”; one-way ANOVA: F(19,180)=4.99, p<0.0001. see Figure 3.3A). These differences serve as another signature with which to compare neural representations. Thus, we also computed separate accuracies for each emotion category in each ROI, and correlated these with the behavioral emotion accuracies. In all ROIs, the accuracy of neural classifications for
different emotions was significantly correlated with the accuracy levels observed in the emotion judgments of the behavioral raters on MTurk (Figure 3A): DMPFC: $r(18)=0.70$, $p=0.001$; MMPFC: $r(18)=0.53$, $p=0.017$; VMPFC: $r(18)=0.47$, $p=0.036$; RTPJ: $r(18)=0.55$, $p=0.012$; LTPJ: $r(18)=0.71$, $p<0.001$; PC: $r(18)=0.46$, $p=0.042$; RSTS: $r(18)=0.65$, $p=0.002$ (see Fig 2b for scatterplot in DMPFC). Thus, the reliable across-emotion accuracy differences observed behaviorally are paralleled in the emotion-specific accuracies of these neural populations (see Fig S2B for neural confusion matrices).

![Figure 3.3](image)

**Figure 3.3.** Classification accuracy broken down by emotion: A. Average classification accuracy for each emotion condition (+/- SEM across exemplars) in behavioral judgments. B. Correlation between behavioral classification accuracies (from A) and neural classification accuracies for each emotion class (based on errors of an SVM trained and tested on DMPFC voxel patterns).

3.2.b RSA

Representational similarity analyses were then used to test specific hypotheses about the structure of the representations in these regions. We generated three competing feature spaces using independent behavioral ratings (Figure 3.4A) and tested which feature space could best capture the neural representation of the 20 emotions. We first analyzed the behavioral data alone, assessing the extent to which emotion categories could be reliably classified based on feature vectors in each of these candidate spaces. Do any of these feature spaces provide a stimulus representation sufficient to match the performance of human subjects in discriminating these
20 emotions (65%)? We found that although all three feature spaces were well above chance level of 5%, the appraisal feature space outperformed the other lower-dimensional spaces (57%, compared to behavioral benchmark of 65%; see Figure 3.4B). Using a paired samples t-test across individual items, we found that the abstract appraisal space performed reliably better than the circumplex space ($t(199)=8.288, p<0.001$) and the basic emotion space ($t(199)=2.176, p=0.031$).

*Figure 3.4.* Competing behavioral feature spaces: Matrix of emotions x average dimension scores for A) the 38-dimensional appraisal space, B) the 6 basic emotion space, and C) the circumplex space. D. Classification of 20 emotions (across stimulus exemplars) using information from each of the 3 competing spaces (+/- SEM across exemplars). Orange dotted line reflects chance (.05); blue dotted line reflects behavioral performance (.65).

These feature spaces were then compared to neural RDMs in each region to identify the space that best accounts for the similarity of the conditions in their neural patterns. In addition to the 3 spaces described above, we tested a model in which condition similarity is defined in terms of similarity of word-frequency vectors (see Experimental Procedures), following the approach of previous attempts to characterize neural semantic representations in terms of word frequencies or word co-occurrences (Mitchell et al., 2008; Murphy, Talukdar, & Mitchell, 2012; Pereira, Detre, & Botvinick, 2011). Does the 38 dimensional appraisal space, which represents emotions in terms of a set of abstract intermediate variables, outperform a raw word-level representation of the stimuli? We also tested three control spaces capturing possible lower-level dimensions of the stimuli—reading ease, syntactic complexity, and rated intensity (confounded with motor response).
For each region, we correlated the RDMs for the competing feature spaces to neural RDMs from individual ROIs (distances of the 20 emotions in their voxel-wise patterns). In the two MPFC subregions, the similarity of conditions in their voxel level patterns was positively correlated with their similarity in the space of 38 appraisal dimensions (at the group-level—DMPFC: kendall’s tau = 0.28; MMPFC: kendall’s tau= 0.21). Moreover, correlations with individual subject neural RDMs (Figure 3.5) revealed a reliable relationship between the neural and model RDMs (DMPFC: M(SEM) kendall’s tau=0.08(0.02), z(19)=3.32 p<0.001; MMPFC: 0.06(0.02), z(19)=2.95 p=0.002). In both of these regions, the correlation with the 38-dimensional space reached the lower bound of the noise ceiling (suggesting that although the average correlations are low, they approach the theoretical maximum given noise in the individual neural RDMs; see Experimental Procedures). Moreover, in both DMPFC and in MMPFC, the neural similarity space was more correlated with the space of 38 appraisal features than with either of the two dimensional spaces: the basic emotions space (DMPFC: 0.08 vs. 0.05, z(19)=3.02, p=0.002; MMPFC: 0.06, vs. 0.03, z(19)=2.31, p=0.021) and the circumplex space (DMPFC: 0.08 vs. 0.06, z(19)=2.84, p=0.005; MMPFC: 0.06 vs. 0.04, z(19)=2.80, p=0.005).
In both regions, the space of abstract appraisal features also outperformed a similarity space defined in terms of word-token frequencies (DMPFC: 0.08 vs. 0.02, z(19)=2.99, p=0.003, MMPFC: 0.06 vs. 0.02, z(19)=2.17, p=0.030), a representation frequently used in fully automated approaches to emotional text classification such as sentiment analysis of reviews or other social media (Pang & Lee, 2004; Tan, Cheng, Wang, & Xu, 2009). To control for lower-level properties of the verbal stimuli, we also compared the neural RDMs to the similarity of stimuli in their reading ease, their syntactic complexity, and rated intensity (confounded with motor response). In both regions, the correlation with the space of 38-appraisals was higher than for reading ease (DMPFC: 0.08 vs. 0.02, z(19)=2.39, p=0.017, MMPFC: 0.06 vs. 0.01, z(19)=2.02, p=0.044), syntactic complexity, and rated intensity.
complexity (DMPFC: 0.08 vs. 0.03, z(19)=2.50, p=0.012, MMPFC: 0.06 vs. 0.02, z(19)=1.98, p=0.048), and intensity (DMPFC: 0.08 vs. 0.02, z(19)=3.21, p=0.001, MMPFC: 0.06 vs. 0.03, z(19)=2.05, p=0.040).

In addition to our a priori ROIs, we conducted the same analyses in the remaining ToM regions (RTPJ, LTPJ, PC, RSTS, and VMPFC): these ROIs were also reliably correlated with the space of 38-appraisals (RTPJ: M(SEM)=0.07(0.02), z(21)=3.59 p<0.001 (see Figure 3.5); see Figure S6 for results from other ToM regions), and no region was reliably more correlated with the basic emotion or circumplex spaces. The 38-dimensional space outperformed competing spaces in all ToM regions except for VMPFC (where the best performing space was the word frequency representation). In VMPFC, RSTS, and RTPJ (but not in PC and LTPJ) the neural-model correlations passed the lower bound of the noise ceiling (Figure S6). However, DMPFC and MMPFC were the only regions in which the high-dimensional space significantly outperformed all other models.

3.2.c Region contributions

We could reliably decode emotion in all of the theory of mind ROIs, and the same 38-dimensional feature space did the best job of capturing the neural similarity space in all regions other than VMPFC. Is the same information represented redundantly across these regions, or is there evidence that these regions contribute differently to the representation of emotions? To address this question, we first compare classification accuracies using single ROIs to the classification accuracy when combining regions across the ToM network. When classifying only valence, a model trained with voxels from all ToM ROIs (M(SEM)=0.581(0.016), t(21)=4.942, p<0.001) performs less well than a model trained only with voxels in DMPFC or MMPFC (58.1% relative to 61% in DMPFC). In contrast, when classifying the full set of 20 emotions, a model trained with voxels from all regions of the network outperforms any of the individual ROIs. Classification using the voxels from all regions of the ToM network (M(SEM)=0.108(0.006), t(21)=9.135, p<0.001; see Figure 3.5) was reliably higher than classification using only voxels in DMPFC (t(38)=2.684, p=0.015), MMPFC (t(38)=2.848, p=0.01), or RTPJ (t(42)=2.773, p=0.011), suggesting that the individual ROIs could contribute non-redundant information.

To further characterize representational differences across the ROIs, we explored whether the regions differ in the particular situation features they represent. Rather than compute separate RDMs for all 38 appraisal
features, we identified a reduced set of 10 features that capture the most unique variance in behavioral ratings across items (see Figure S5; Experimental Procedures). We then computed the RDMs for this 10-dimensional space, and also for each of the 10 features individually, and correlated each with the neural RDMs in different regions. Thus in each subject, we obtained neural-feature correlations for each of the 10 features in each ROI. By testing for feature x ROI interactions across subjects, we can thus test for differences in the feature representations across ROIs. We focus in particular on comparing MPFC and RTPJ, as these regions have been proposed to be involved in distinct aspects of mental state reasoning (affective and epistemic respectively; see Koster-Hale et al., in review). The neural RDMs in DMPFC, MMPFC, and RTPJ were reliably correlated with the RDM of the 10-feature space (DMPFC: M(SEM)=0.08(0.02), z(19)=3.21 p=0.001; MMPFC: M(SEM)=0.05(0.02), z(19)=2.61 p=0.004; RTPJ: M(SEM)=0.06(0.01), z(21)=3.55 p<0.001), and this smaller set of features appears to capture much of the representational structure of the initial 38-d space (Figure 3.6; see Figure S7 for results from secondary ROIs). A repeated-measures ANOVA on the neural-model correlations for each feature (with ROI and feature as within-subjects factors) revealed a significant ROI x feature interaction for the comparison of DMPFC and RTPJ (F(9,171)=2.06, p=0.036) but not between MMPFC and RTPJ (F(9,171)=1.036, p=0.414).

The differential feature representations between DMPFC and RTPJ suggest that although multiple ToM regions are involved in the attribution of emotion, some of these regions may contribute unique information to the final representational space that governs behavior. For example, exploratory analyses reveal that the correlation with the “self cause” feature (“Was this situation caused by <character> herself or by someone/something else?”) is reliably higher in DMPFC than in RTPJ, where as the “distant past” feature (“Did this situation involve events from <character>’s distant past?”) is more correlated with the RDM in RTPJ than in DMPFC.
3.3 DISCUSSION

The ability to predict and infer the emotional states of others is central to our species’ unique social and cooperative behaviors (Tomasello, Carpenter, Call, Behne, & Moll, 2005). In the present research, we provide evidence that neural patterns in regions involved in mental state reasoning contain information relating to the emotional states of others. Moreover, we provide quantitative insight into the underlying representational structure that supports this inferential ability.

3.3.a The structure of emotion knowledge
Decades of research in the science of emotion have aimed to characterize emotions in terms some low-dimensional space of basic affective primitives (Ekman, 1992; Posner et al., 2009b; Russell, 1980). Behaviorally, we find that a space of 38 abstract event features, inspired by work in appraisal theory (Scherer & Meuleman, 2013), reliably outperforms these simpler spaces in discriminating the 20 different emotions in our stimuli. While affective dimensions that make up the circumplex model (Russell, 1980) or basic emotions theory (Ekman, 1992; Ekman & Rosenberg, 1997) may capture the range of emotions we express and perceive with overt expressions, a higher-dimensional space is needed to encode the range attributions elicited by short verbal descriptions of events.

Interestingly, a model using the 38-dimensional space still falls short of human behavioral performance when labeling stimuli (57% versus 65% accurate), indicating that this collection of features does not adequately capture our intuitive emotion knowledge. There are at least two plausible reasons for this inadequacy. First, the 38 dimensions used in the present study were derived from prior literature on emotion appraisal without subsequent optimization. This list may therefore contain redundant or uninformative features, and some additional features are likely necessary.

A second, more fundamental limitation is that this approach aims to encode human emotion knowledge in terms of flat feature vectors (i.e. lists of appraisal checks applied to each stimulus). While this feature-based approach has been productive in other domains of perception and cognition (Freiwald, Tsao, & Livingstone, 2009; Just, Cherkassky, Aryal, & Mitchell, 2010; Koster-Hale, Bedny, & Saxe, 2014; Koster-Hale, Saxe, Dungan, & Young, 2013; Sudre et al., 2012a) and proved useful in the present paradigm, it is unlikely that representations in a domain of high-level cognition such as theory of mind can be reduced to operations over lists of associated features (Laurence & Margolis, 1999; Murphy & Medin, 1985). For example, emotions are caused by events that unfold over time: the emotion attributed depends critically on the temporal and causal order of the different elements of the event (e.g. eating a whole cake and then swearing to keep to your diet; versus swearing to keep your diet and then eating a whole cake). To capture the richly causal and compositional nature of the representations involved in emotion inference (Ortony, 1990), future research will need to move beyond a
feature-based approach, incorporating structured, generative knowledge representations from other areas of cognitive science (Baker, Saxe, & Tenenbaum, 2009; Tenenbaum, Kemp, Griffiths, & Goodman, 2011).

Nonetheless, present research makes important advances in our understanding of emotion inference. While constructivist theories of emotion have long acknowledged that attribution depends on emotion-specific conceptual knowledge (Barrett, 2006a; Brosch, 2013; Clore & Ortony, 2013; Lindquist, 2013; Lindquist & Barrett, 2008; Lindquist, Gendron, Barrett, & Dickerson, 2014), the content and structure of that knowledge has remained unclear. Here, we provide an initial sketch of specific features that might structure human emotion concepts, and provide a framework for evaluating competing models of this knowledge.

### 3.3.b Neural representations of others’ emotions

Consistent with previous reports (Peelen et al., 2010; Skerry & Saxe, 2014), the present results suggest that neural representations in MPFC contain information about attributed emotions. Whereas prior studies focused on coarse distinctions (e.g. valence), we are able to classify a set of nuanced emotions at above chance levels, suggesting that emotion representations in this region are relatively fine-grained. Moreover, by expanding to a rich space of eliciting situations, we are able to decode attributed emotions in all regions of the ToM network. Using a whole brain searchlight, we find that although emotion information is present in many regions, this information is largely restricted to regions involved in theory of mind (particularly MPFC, RTPJ, and LTPJ). When combining information from voxels across all ToM regions, we were able to decode the emotion label of a stimulus with ~10% accuracy.

Although these classifications were reliably above chance (5%), they are far from reaching the accuracy observed in behavior (65%). This discrepancy between neural and behavioral classification could arise because the population code in these regions is insufficient to explain the behavior, or because single trial estimates of fMRI data provide a noisy, highly blurred measurement of the underlying neural code. However, across different emotions, there were reliable correlations in the average accuracy of the neural populations and of independent behavioral raters, providing support for the role of these regions in emotion attribution behaviors.

We then further probed the underlying representational structure that supports successful emotion
discrimination. The previous literature (Chikazoe et al., 2014; Peelen et al., 2010; Skerry & Saxe, 2014) is consistent with the possibility that MPFC codes a limited space of affective dimensions such as valence and/or arousal. Moreover, even in our MPVA analyses, a region could support 20-way classification at above-chance levels by coding only a single dimension or feature that varies across emotions. Using RSA, we find that brain regions selective for theory of mind not only contain information about attributed emotions, they are also best captured by the high-dimensional space of event features.

In all but one of the ToM regions, the similarity of emotion conditions in their voxel response patterns was most correlated with the similarity of the emotions in the space of 38 appraisals. This result suggests that the neural code in these regions does not reduce to a simpler set of distinctions such as valence and arousal, and provides novel insight into the granularity of the emotion representations in MPFC and other ToM regions. Together, the behavioral and neural data suggest that human emotion attribution is organized around abstract features of the causal context in which different emotions occur, rather than the sorts of affective primitives that have dominated prior research.

A challenge for future work will be characterizing the scope and specificity of the neural representations in the observed regions. One possibility is that these neural populations contain representations specific to attributed emotion, and that these attributed states are coded within a space of emotion-relevant causal features. Alternatively, there could be neural populations that contain information about emotion-relevant features, but in the form of domain-general semantic representations. These event representations might serve as intermediate features in the service of diverse inferential processes beyond emotion attribution. Ultimately, successful emotion inference depends on a rich body of general world knowledge, and neural populations specific to social cognition must interface with more general-purpose semantic processing mechanisms. Characterizing information flow within and between these different networks will be an important avenue for future research.

3.3.c Relation to prior research

In the present research, we provide a first attempt to characterize the feature space that governs emotion representation in the human brain. To do so, we draw heavily on methods and ideas that have been fruitful in
recent research on visual object recognition and object semantics, where researchers have tested a range of high-level and low-level features that could capture neural similarity of different objects (Baldassi et al., 2013; Carlson, Simmons, Kriegeskorte, & Slevc, 2013; Kriegeskorte, Mur, & Bandettini, 2008; Liu et al., 2013; Mitchell et al., 2008; Palatucci, Pomerleau, Hinton, & Mitchell, 2009; Pereira et al., 2011; Rust & DiCarlo, 2010). In one study, Mitchell and colleagues (Mitchell et al., 2008) coded object words in terms of their co-occurrence with a set of 25 verbs hypothesized to pick out relevant semantic dimensions (e.g. “manipulate”, “taste”), and found that this representation was sufficient to support above chance neural classification of untrained stimuli. Further analyses of these data show that a corpus-based co-occurrence space is outperformed by a space derived from behavioral ratings on a set of a priori object properties (e.g. is it alive?) (Palatucci et al., 2009; Sudre et al., 2012b). The present research is most similar to this second approach, relying on behavioral ratings of a set of hypothesized event features. We show that it is possible to generate candidate representational spaces for domains of high-level cognition such as emotion inference, and to use these spaces to characterize patterns of activity in theory of mind brain regions.

With this approach, we hope to move beyond identifying regions that contain information about emotion attributions, and gain insight into the intermediate stages and corresponding features used to construct these attributions. In the study of object representation, researchers have made headway in understanding differences across regions and temporal stages (Carlson et al., 2013; Isik, Meyers, Leibo, & Poggio, 2013; Rust & DiCarlo, 2010); representational similarity analysis in particular has provided a flexible framework for comparing the structure of the representations in different regions along the ventral pathway (Cichy, Pantazis, & Oliva, 2014; Leeds, Seibert, Pyles, & Tarr, 2013). Interestingly, the present results provide preliminary evidence that theory of mind regions differ in their contributions to emotion inference. When classifying the 20 emotions, we find a reliable advantage to using voxels from the whole network, compared to any region in isolation. Moreover, we observed region-by-feature interactions in the RSA analyses, suggesting that the regions differ in the specific appraisal features that dominate their response. Further work is needed to characterize the precise computational roles of these regions and how they interact with other networks to form a processing stream.
As has been the case in research on object representation, we assume that future studies of emotion attribution will yield feature spaces that outperform the 38-dimensional space explored here. Future work might aim to not only better fit the neural data, but also to build computational models capable of extracting the relevant intermediate features directly. Many early approaches to modeling neural object representations involved hand-picked feature spaces (e.g. 25 chosen verbs) (Jelodar et al., 2010; Mitchell et al., 2008) and often manual coding of stimuli within those spaces (Carlson et al., 2013; Chang, Mitchell, & Just, 2011; Huth, Nishimoto, Vu, & Gallant, 2012; Palatucci et al., 2009). However, recent research has yielded computational models that can be applied to raw stimuli (i.e. images) and achieve high quantitative fit to neural patterns (Yamins et al., 2014); even relevant feature spaces themselves can be discovered in a bottom-up manner (Devereux, Kelly, & Korhonen, 2010; Murphy et al., 2012; Stansbury, Naselaris, & Gallant, 2013). In our study, candidate features were selected based on prior domain knowledge, and the stimuli required manual annotation into these feature spaces (MTurk ratings). Future research in this area should ideally identify new sets of optimized features (either event features or some other candidate basis), and new ways to infer these features from text alone, removing the need for a human subject or experimenter in the loop.

Despite these important open questions, the present data provide novel insight into the representations underlying human emotion inference and the neural populations that support them. Together, the results suggest that our knowledge of others’ emotions is abstract and high-dimensional, that brain regions associated with emotion perception and inference contain information about relatively fine-grained emotional distinctions, and that the neural representations in these regions not reducible to more primitive affective dimensions such as valence and arousal.

3.4 EXPERIMENTAL PROCEDURES

Further details on experimental procedures (e.g. ROI selection and univariate analyses) are provided in the Supplemental Experimental Procedures.

3.4.a Stimuli
All experiments used a set of 200 verbal stimuli (2-3 sentences; M(SEM)= 50.68(0.28) words; see Table 1) describing a character experiencing one of 20 different emotions. In each item, the emotion was conveyed via a description of an emotion-eliciting event, without any labeling or description of the character’s reaction.

3.4.b Behavioral attributions

To verify that subjects make reliable attributions of the emotions conveyed in the 200 stimuli, subjects on Amazon Mechanical Turk (MTurk, N=139) were asked to choose which of the 20 emotions best described the character’s emotional state (see Supplemental Experimental Procedures). Predicted emotions for each stimulus were used to compute an overall accuracy level (relative to the intended emotion for each stimulus; see Figure 3.4A), as well as a confusion matrix (the proportion of time each intended emotion was labeled as each of the emotion categories; see Figure S4A).

3.4.c Behavioral feature ratings

A separate set of MTurk subjects (N=250) provided ratings (1-10 scale) for each of the stimuli on each of the features of the three competing feature spaces. A given subject rated stimuli on either features from the 38-dimensional appraisal space (e.g. “Did someone cause this situation intentionally or did it occur by accident?” 1= caused accidentally, 5=neutral/not applicable, 10=caused intentionally; see Supplemental Table), or dimensions corresponding to the basic emotion space (e.g. “Was <character> happy in this situation?” 1= not at all happy, 5=somewhat/not applicable, 10=very happy) and the circumplex space (e.g. “Did <character> find this situation to be positive or negative?” 1=clearly negative, 5=neutral/not applicable, 10=clearly positive).

3.4.d Feature-based classification of behavioral data

To test whether any of the 3 candidate spaces (basic emotion, circumplex, and 38 appraisals) capture the full range of attributed emotions, we created an item-by-feature matrix for each possible space, and tested whether a model trained on these features could classify the 20 distinct emotions. Specifically, we trained a linear SVM (SVC, one vs. one implementation) on emotion-labeled feature vectors for a subset of items, and tested whether the classifier could generate the appropriate label for a different set of items. Thus, we test whether each feature
space provides a basis for emotion discrimination that generalizes across the different exemplars\textsuperscript{10}. We conducted this procedure iteratively (n=1000), splitting items into 100 training and test exemplars (5 items for each emotion condition), and computed the average cross-item classification accuracy for each feature space, to compare to the behavioral benchmark (65%). We also computed the accuracy separately for each item, and tested for reliable differences between the three feature spaces using a t-test across items.

3.4.e FMRI participants

22 right-handed adults ages 18-40 (M\textsubscript{age} = 25.39, STD\textsubscript{age}=5.43; 13 female) participated in the study. All participants had normal/corrected-to-normal vision and no history of neurological disorders. Participants gave written, informed consent in accordance with the requirements of the MIT institutional review board. We collected behavioral measures of social-cognitive ability from each participant (see Supplementary Experimental Procedures).

3.4.f FMRI tasks

\textit{Theory of mind localizer:} Subjects were presented with short textual scenarios that required inferences about mental states (Belief condition) or physical representations such as a map or photo (Photo condition; (Dodell-Feder et al., 2011; Saxe & Kanwisher, 2003) (stimuli available at http://saxelab.mit.edu/superloc.php). Scenarios were presented for 10s, followed by a true or false question (4s) about either the representation (Belief or Photo) or the reality of the situation. Each run (4.53min) consisted of 10 trials separated by 12s inter-stimulus intervals, and 12s blocks of fixation were included at the beginning and end of each run. 1-2 runs were presented to each participant, with the order of stimulus type (Belief or Photo) and correct answer (True or False) counterbalanced within and across runs.

\textit{Emotion Attribution task:} In the Emotion Attribution task, subjects viewed the 200 emotion

\begin{footnotesize}
\textsuperscript{10} Note, because the set of 20 emotions includes some basic emotions (disgust, surprised), the comparison of the 38-appraisal spaces to the space of basic emotions is rather conservative. That is, the basic emotion space actually contains as features some of the labels we are trying to predict (i.e. to predict emotion categories like “disgusted”, “surprised”, and “terrified”, the basic emotion space uses ratings of the extent to which the event elicits “disgusted”, “surprised”, “afraid” emotions). If the 38-dimensional space outperforms the basic emotion space despite being at such a disadvantage, this would provide particularly compelling evidence for the role of causal context in structuring our representations of others’ emotions.
\end{footnotesize}
stimuli, as well as a set of 10 stories describing physical pain (Bruneau et al., 2012b). The experiment consisted of 10 runs (7.37 min/run), each containing 1 exemplar for each of the 21 conditions (20 emotions plus 1 pain stimulus). Each story was presented at fixation for 13s, followed by a 2s window during which subjects made a behavioral response. Subjects were instructed to press a button to indicate the intensity of the character’s experience (1 to 4, neutral to extreme), which focused subjects’ attention on the character’s emotional state, but ensured that behavioral responses (intensity) were orthogonal to discriminations of interest. The stories were presented in a jittered, event-related design, with a central fixation cross presented between trials at a variable inter-stimulus interval of 3-5-7 seconds. The order of conditions was counterbalanced across runs and participants, and the order of individual stories for each condition was randomized.

3.4.g FMRI acquisition

Data were acquired on a 3T Siemens Tim Trio scanner in the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT, using a Siemens 32-channel phased array head coil. We collected a high-resolution (1mm isotropic) T-1 weighted MPRAGE anatomical scan, followed by functional images acquired with a gradient-echo EPI sequence sensitive to blood-oxygen-dependent (BOLD) contrast (repetition time [TR] = 2s, echo time [TE] = 30ms, flip angle = 90°, voxel size 3x3x3mm, matrix 64x64, 32 axial slices).

3.4.h FMRI analyses

Preprocessing: MRI data were preprocessed using SPM8 (www.fil.ion.ucl.ac.uk/spm/software/spm8/), freesurfer (for skull-stripping; http://surfer.nmr.mgh.harvard.edu/), and in-house code. SPM was used to motion correct each subject’s data via rigid rotation and translation about the 6 orthogonal axes of motion, to register the functional data to the subject’s high-resolution anatomical image, to normalize the data onto a common brain space (MNI), and to smooth images with a Gaussian filter (FWHM=5mm).

MVPA Classification Analyses: We first aimed to replicate previous valence decoding in MPFC (Skerry and Saxe, 2014) by choosing subset of conditions that most closely match the happy versus sad emotions used in that study (“Excited”, “Joyful”, “Proud” vs. “Devastated”, “Disappointed”, “Annoyed”) and testing
whether voxel patterns in MFPC could reliably classify the valence of these stimuli. We then tested whether voxel patterns in MFPC or other ToM regions could reliably classify the set of 20 emotions. Code for multi-voxel pattern classification was developed in Python using the publicly available PyMVPA toolbox (http://www.pymvpa.org/).

We conducted MVPA within ROIs that were functionally defined based on individual subject localizer scans. High-pass filtering (128 Hz) was conducted on each run, with linear detrending across the whole time-course. A timepoint was excluded if it was a global intensity outlier (> 3 SD above the mean intensity) or contained a large movement (> 2mm scan-to-scan). The data were temporally compressed to generate a voxel-wise average for each individual trial, and these single trial summaries were used for training and testing. The individual trial patterns were calculated by averaging the preprocessed bold images for the 12TR duration of the trial, offset by 5 TRs to account for the HRF and lag in relevant context (offset and duration selected based on subset of time-course in which the response magnitude differed between pain and emotion stimuli). Rest timepoints were removed and the trial summaries were concatenated. The pattern for each trial was then z-scored relative to the mean across all trial responses in that voxel.

The data were classified using a support vector machine; this classifier uses condition-labeled training data to learn a weight for each voxel, and subsequent stimuli (validation data not used for model training) can then be assigned to one of two classes based on a weighted linear combination of the responses in each voxel. For the 20-way discrimination, multi-class classification was conducted with a one-vs-one method (Knerr, Personnaz, & Dreyfus, 1990), yielding a single condition prediction for each trial. We used a fixed regularization parameter (C=1) and restricted ourselves to linearly decodable signal under the assumption that a linear kernel implements a plausible readout mechanism for downstream neurons (Hung, Kreiman, Poggio, & DiCarlo, 2005; Seung & Sompolinsky, 1993; Shamir & Sompolinsky, 2006).

The data were partitioned into 10 run-based folds and the classifier was trained iteratively on all runs but one, and tested on the remaining run. Classification accuracy was averaged across folds to yield a single score for each subject in the ROI. A one-sample t-test was then performed over these individual accuracies, comparing to
chance classification (.5 for positive versus negative, and .05 for the 20-way emotion classification; all t-tests on classification accuracies were one-tailed). We also performed this analysis using the full ToM network, where each subject’s ROI was the union of his/her individually localized ROIs.

**Whole brain searchlight classification:** The searchlight procedure was identical to the ROI-based procedure except that the classifier was applied to voxels within local spheres rather than individually localized ROIs. For each voxel in a gray matter mask, we defined a sphere containing all voxels within a 3-voxel radius (123 voxels) of the center voxel. Classification was then performed on each cross-validation fold, and the average classification accuracy for each sphere was assigned to its central voxel, yielding a single accuracy image for each subject for a given discrimination. A one-sample t-test over subjects’ accuracy maps (comparing accuracy in each voxel to chance—0.05) yielded a group t-map, which was assessed at a p<.05 (K>25), FWE corrected (based on SPM’s implementation of Gaussian Random Fields).

**Representational Similarity Analyses:** To create representational dissimilarity matrices (RDMs) for the competing feature spaces, we first averaged the feature vectors for each emotion condition (across stimuli), yielding the emotion-by-feature matrices shown in Figure 3.4. For each matrix, we then computed the Euclidean distance of feature vectors for each pair of emotions. We conducted this analysis iteratively (n=1000) across split halves of the data (5 items per condition in each half), such that the self-distances along the diagonal are meaningful.

In addition to the three candidate feature spaces described above (circumplex model, basic emotions, and abstract appraisals), we generated an additional space defined in terms of the similarity in word occurrences across stimuli. Features vectors were created based on frequencies of individual words from the stimuli, excluding English stop words and stripping common morphological endings. We used a term frequency-inverse document frequency (tf-idf) vectorizer such that the exemplar value for each word increases with the frequency of the word in the exemplar, but decreases with the frequency of the word in the full stimulus set. This type of flat, unordered word-level representation is popular in existing machine learning approaches to sentiment analysis/emotion classification (Pang & Lee, 2004; Tan et al., 2009). Finally, we computed several additional control spaces to
confirm that neural RDMs could not be explained in terms of lower-level properties of the stimuli: reading ease, syntactic complexity, and behavioral ratings of intensity (see Supplemental Procedures).

Neural RDMs were computed separately for each region in each subject. These were computed with a procedure analogous to that described for feature space RDMs, except that the features were voxel-wise neural responses rather than the behavioral feature ratings. We averaged voxel response vectors for each condition separately split halves of the data, yielding two condition-by-voxel matrices (for even and odd runs). We then computed similarity of the conditions in terms of Euclidean distance of the voxel patterns across runs, yielding a RDM for each region (again this is done across even and odd subsets so that the diagonal is interpretable). Each neural RDM was normalized by subtracting its minimum value and dividing by the range, yielding a matrix with distances ranging from 0 to 1. This procedure was conducted separately for each individual subject, and individual subject neural RDMs for each region were averaged to generate a group RDM for the region.

To compare neural and model similarity spaces, we then computed the rank correlation (kendall’s tau-a) between the neural and model RDMS for each region. The group neural RDM will be least noisy, and therefore provide the best estimate of the relationship between the true neural RDM and each of the model spaces. However, to assess the reliability of the neural-model relationships, we compute the neural-model correlations separately for each subject, and perform a Wilcoxon test comparing the individual subject correlations to chance (average kendall’s tau = 0). We also compare the fit of different models by conducting a one-tailed Wilcoxon signed-rank test on the correlations for different pairs of models.

We compare these neural-model correlations to a behavioral benchmark (dotted line in Figures 3.5 and 3.6) was defined as the correlation of the neural RDM with an RDM computed from the confusion matrix of independent behavioral raters. From the behavioral classification study, we have scores representing the frequency with which a given emotion is misclassified as another emotion. We then computed Euclidean distances of the conditions with this matrix to yield a RDM capturing the similarity space of the emotion conditions (though because behavioral performance was relatively high, the confusion matrix is sparse and may underestimate the similarity structure of the emotions). We then test whether any of our intermediate feature
spaces (38-dimensions, basic emotions, circumplex model, etc.) meet or surpass this benchmark correlation. Given that our analyses depends on neural-model correlations computed in individual subjects, we wished to assess the extent to which our correlations are limited by noise in the individual subject neural RDMs. We estimated a noise ceiling for this analyses approach, on the assumption that individual neural RDMs can be no more correlated with a model than they are with the true neural RDM (Nili et al., 2014). We computed the correlation of each individual RDM to the mean RDM of the full group, which potentially overestimates the reliability of the individual RDMs since the individual subject’s data is included in the group. We then used a leave-one-out procedure to correlate the individual RDMs to the mean of the group RDM excluding that subject; this potentially underestimates the reliability of the individual neural RDMs with the true neural representation, since the group mean is an average of a small sample of subjects. These two values serve as upper and lower bounds on the neural-model correlations we can expect to observe with this analysis approach.

Region analysis—comparisons of individual features: Using a set of 10 features that explain the most unique variance across stimuli (see Supplemental Procedures and Data), we created a reduced 10-dimensional space and subject it to the same behavioral and neural analyses described above. Specifically, we compute RDMs for the 10-feature space, and for each feature in isolation, and correlate these with the neural RDMs in each region. To assess whether regions differ in the individual features they represent, we conducted repeated measures ANOVA on the correlations (kendall’s tau) between neural and the feature RDMs, with ROI and feature as within-subject factors.

3.5 SUPPLEMENTAL INFORMATION

3.5.1 Supplemental Experimental Procedures

Supplemental Table. 38 features used to construct the abstract appraisal space (RSA analysis)

<table>
<thead>
<tr>
<th>Feature Name</th>
<th>Feature Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>expectedness</td>
<td>Did &lt;character&gt; expect this situation to occur?</td>
</tr>
<tr>
<td>pleasantness</td>
<td>Did the situation involve a hedonically positive or pleasant experience for &lt;character&gt;?</td>
</tr>
<tr>
<td>goal consistency</td>
<td>Was the situation consistent or inconsistent with &lt;character&gt;’s goals, needs,</td>
</tr>
<tr>
<td>(Continued)</td>
<td>or desires?</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>fairness</td>
<td>Was this situation fair or unfair for &lt;character&gt;?</td>
</tr>
<tr>
<td>caused by agent</td>
<td>Was this situation caused by a person or some other external force (e.g. randomness)?</td>
</tr>
<tr>
<td>intentional action</td>
<td>Did someone cause this situation intentionally or did it occur by accident?</td>
</tr>
<tr>
<td>caused by self</td>
<td>Was this situation caused by &lt;character&gt; herself or by someone/something else?</td>
</tr>
<tr>
<td>involved close others</td>
<td>Did this situation involve people that &lt;character&gt; felt close to?</td>
</tr>
<tr>
<td>control</td>
<td>Were the events in this situation primarily within &lt;character&gt;’s control?</td>
</tr>
<tr>
<td>altering</td>
<td>Did &lt;character&gt; think she had the power to alter the situation in the future?</td>
</tr>
<tr>
<td>morality</td>
<td>Did this situation involve people behaving in a way that would be considered proper or moral?</td>
</tr>
<tr>
<td>self esteem</td>
<td>Did this situation affect &lt;character&gt;’s self-esteem or opinion of herself?</td>
</tr>
<tr>
<td>suddenness</td>
<td>Did this situation occur suddenly/out of the blue?</td>
</tr>
<tr>
<td>familiarity</td>
<td>Was this situation a familiar event/situation for &lt;character&gt;?</td>
</tr>
<tr>
<td>future</td>
<td>Did &lt;character&gt;’s emotion involve an event that would or might occur in the future?</td>
</tr>
<tr>
<td>distant past</td>
<td>Did this situation involve events from &lt;character&gt;’s distant past?</td>
</tr>
<tr>
<td>already occurred</td>
<td>Was &lt;character&gt;’s emotion based on an something that had already occurred?</td>
</tr>
<tr>
<td>certainty</td>
<td>Did &lt;character&gt; feel certain about the situation/outcome?</td>
</tr>
<tr>
<td>repetition</td>
<td>Did &lt;character&gt; think the situation was likely to occur again?</td>
</tr>
<tr>
<td>coping</td>
<td>Did &lt;character&gt; think she could cope with/handle the situation?</td>
</tr>
<tr>
<td>mental states</td>
<td>Was &lt;character&gt;’s emotion related to the mental states (e.g. beliefs, attitudes) of other people?</td>
</tr>
<tr>
<td>others knowledge</td>
<td>Did people other than &lt;character&gt; know about the situation that occurred?</td>
</tr>
<tr>
<td>bodily disease</td>
<td>Did the situation involve events relating to the physical body?</td>
</tr>
<tr>
<td>other people</td>
<td>Was &lt;character&gt; interacting with people in this situation?</td>
</tr>
<tr>
<td>self relevance</td>
<td>Was there a lot at stake for &lt;character&gt; in this situation? Did the events have high-relevance for &lt;character&gt;’s life?</td>
</tr>
<tr>
<td>freedom</td>
<td>Was &lt;character&gt; free to act or behave however she wanted in this situation?</td>
</tr>
<tr>
<td>pressure</td>
<td>Was &lt;character&gt; under a lot of pressure in this situation?</td>
</tr>
<tr>
<td>consequences</td>
<td>Was &lt;character&gt;’s situation an isolated incident, or did it have long-term consequences?</td>
</tr>
<tr>
<td>danger</td>
<td>Was &lt;character&gt; in physical danger in this situation?</td>
</tr>
<tr>
<td>self involvement</td>
<td>Did the situation describe an outcome directly involving &lt;character&gt; herself or primarily involving other people?</td>
</tr>
<tr>
<td>remembering</td>
<td>Was the situation something that &lt;character&gt; is likely to remember in the future?</td>
</tr>
<tr>
<td>self consistency</td>
<td>Did the situation involve events consistent with &lt;character&gt;’s personality or self-concept?</td>
</tr>
<tr>
<td>relationship influence</td>
<td>Did this situation affect &lt;character&gt;’s relationships with other people?</td>
</tr>
<tr>
<td>agent vs. situation</td>
<td>Was this event primarily a reflection of &lt;character&gt; (e.g. her personality, her abilities) or a reflection of the surrounding situation?</td>
</tr>
</tbody>
</table>
**Behavioral data acquisition for Classification Analysis:** Behavioral data were collected on Amazon’s Mechanical Turk. To obtain behavioral classification for each of the 200 stimuli, a set of subjects were presented with a single story on each trial and asked to choose which emotion (one of the 20 emotion categories, plus Neutral) best described the emotional state of the character. In addition to the stimuli from the 20 emotion categories, subjects were presented two stories in which the character was explicitly described as feeling neutral, which was used as an attention/quality check. We obtained judgments from 172 subjects (96 female; M(SEM) age = 33.326(0.873), and reduced the sample to 139 subjects who passed the quality check questions (by rating the neutral stories as neutral).

**Behavioral data acquisition for RSA analysis:** To construct RDMs, a separate set of subjects was used to obtain ratings of the stimuli within each of our feature spaces. MTurk subjects (n=250) were presented with a single stimulus item and used a 1 to 10 scale to rate the extent to which the event contained each feature in a given space. As an attention check, subjects were asked to rate the extent to which the story involved the character named in the story (subjects were excluded if their average response to this item was <7, and individual items were excluded if the response on this question was <5). 22 subjects were excluded for failing the attention check, leaving 238 subjects (108 female; M(SEM) age = 34.47(0.77)) and an average of 15.4 responses for each of the 200 items). Subjects were allowed to rate more than one stimulus, and a given subject rated stimuli either on features from the 38-dimensional appraisal space (e.g. “Did someone cause this situation intentionally or did it occur by accident?”), or on dimensions corresponding to the basic emotion space (e.g. “What <character> happy in this situation”) and the circumplex space (“Did <character> find this situation to be positive or negative?”).

**Region of interest selection:** To define individual ToM ROIs, we used hypothesis spaces for
bilateral TPJ, right STS, PC, ventral, dorsal and middle subregions of MPFC, which were derived from previous random effects analyses with this task (see Figure S1 for hypothesis spaces). The task was modeled as a 14s boxcar (the full length of the story and question period) convolved with a standard hemodynamic response function, and a general linear model (implemented in SPM8) was used to estimate beta values for Belief trials and Photo trials. We conducted high-pass filtering at 128hz, normalized the global mean signal, and included nuisance covariates to remove effects of run. For each subject, we used a t-test implemented in SPM8 to generate a map of t values for the contrast of Belief>Photo and identified the peak t value within the hypothesis space. An individual subject’s ROI was defined as the cluster of contiguous suprathreshold voxels (minimum k=10) within a 9mm sphere surrounding this peak. If no cluster was found at p<0.001, we repeated this procedure at p<0.01 and p<.05 (see Figure S1). We masked each ROI by its hypothesis space—defined to be mutually exclusive—such that there was no overlap in the voxels contained in each functionally defined ROI. An ROI for a given subject was required to have at least 20 voxels to be included in multivariate analyses.

**Univariate analyses of mental state selectivity in ToM ROIs:** To confirm the selectivity of the individually localized ROIs, we compared the average BOLD response to the emotion stimuli to the average response to non-mental stories describing events in which a character experienced physical pain. This task was modeled as a boxcar (the full length of the story and response period) convolved with a standard HRF. We conducted high-pass filtering at 128hz, normalized the global mean signal, and included nuisance covariates to remove effects of run. We computed for each ROI the average beta value across voxels for each condition, and then averaged the beta values for the different emotion conditions to compare to the response to pain stimuli. In each ROI, we conducted a paired sample t-test comparing the beta values for emotion and pain conditions (see Figure S1).

**Percent signal change in ToM ROIs:** To visualize the univariate response to all conditions in these regions (see Figure S1), we computed the percent signal change (PSC) relative to baseline for each of the 21 conditions (20 emotions, plus physical pain). Baseline response for each ROI was computed as the average BOLD response at all rest time points, excluding the first 4s after stimulus offset for each trial. The PSC relative to
baseline was calculated for each time point in each condition, averaging across all voxels in the ROI and across all trials in the condition, where PSC (at time t) = (average BOLD for condition at time t – average BOLD for fixation)/average BOLD magnitude for fixation. We plot this event-related average to visualize the average BOLD response at each time point after the trial onset for each condition.

**Construction of control feature spaces for RSA analysis:** To control for possible confounding sources of variation in our stimulus set, we computed three control spaces: reading ease, syntactic complexity, and intensity as rated by subjects in the scanner (confounded with motor response). The similarity spaces for reading ease and syntactic complexity were both derived from features extracted using CohMetrix (www.cohmetrix.com). Reading ease was made up of each exemplar’s scores on Flesch Reading Ease (measuring average sentence length and number of syllables per word, where higher scores indicates easier text and increased readability) and Standardized Cohesion (measuring extent to which words overlap across the text; text with low referential cohesion is usually more difficult to process as there are fewer repetitions connecting ideas across the text). Syntactic complexity was defined in terms of Negation (measuring the number of negative expressions in the text, such as no, not, un-, without), Noun Phrase Modification (measuring the average number of modifiers, such as adjectives, adverbs, and determiners, per noun phrase), and Left-Embeddedness (measuring the average number of words before the main verb in each sentence). Finally, we computed the similarity of each stimulus in terms of its average intensity, derived from the in scanner behavioral judgments (1=neutral/low intensity, 4=extreme/high intensity). This RDM allows us to control for neural patterns reflecting possible differences in motor responses across emotion conditions.

**Feature selection for region comparison analysis:** To characterize the representation of specific appraisal features, we identified a reduced set of features that capture unique variance across stimuli (i.e. to eliminate redundant features or features that do not reliably vary across the stimuli). Computing pairwise correlations between each of the features (see Figure S4), we observe substantial correlations amongst features, suggesting that a smaller feature space may be sufficient. While many approaches to dimensionality reduction involve transforming data into linear combinations of the initial features, we wished to maintain the
interpretability of our semantically meaningful features, and therefore used a forward step-wise regression procedure. On each of 38 iterations, we computed a separate regression fitting each of the available appraisal dimensions to the data (original 200 stimuli x 38 dimension matrix), yielding a separate $R^2$ value for each dimension. We then identified the appraisal dimension with the highest $R^2$ for that iteration and added it to an ordered list of appraisals. We then performed a regression on the initial data matrix using that dimension and the previously selected appraisals, with the $R^2$ characterizing the variance explained in the initial data matrix that can by this reduced set of features. The residuals of this regression (observed-predicted) served as the data matrix for the subsequent iteration. Thus, this procedure selects appraisal features that explain unique variance in the data. The first 10 appraisal features extracted with this method together capture 75.95% of the variance in the full input matrix, and are used as a reduced space. With the behavioral data, we perform 20-way classification using a feature vector in this 10-dimensional space, and using each of the 10 features in isolation (see Figure 3.6).

**Analysis of RSA time-courses for region comparison:** We also explored the temporal profile of representation in each region (see Cichy, Pantazis, & Oliva, 2014) by computing RSA time-courses for each region. To do so, we separately analyzed a series of overlapping 2 TR (4 sec) windows, with onsets ranging from 0 to 11 TRs post stimulus presentation. For each window, we conducted the RSA analyses described above (compute neural RDM within that temporal window, and compute kendall’s tau between the neural RDM and the model RDM for that time period). We can then plot these neural-model correlations over time to identify differences in the time-course of similarities across different region.

**Behavioral individual difference measures:** To assess the relevance of emotion-specific neural patterns to emotional and social competence, we collected behavioral measures of empathy and emotion recognition abilities and sought to relate these to individual differences in neural classification accuracy. The AQ (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001) and the Interpersonal Reactivity Index (IRI, Davis, 1983) were completed via online Qualtrics surveys (www.qualtrics.com), yielding for each participant a single AQ score, and separate IRI scores for Empathic Concern (EC), Fantasy (FS), Personal Distress (PD), and Perspective Taking (PT). Participants also completed an Empathic Accuracy task (Zaki, Bolger, & Ochsner,
2008) in which they made continuous ratings of a target’s emotional state over the course of 16 trials. Empathic accuracy was operationalized as the correlation between subjects’ emotion ratings and ratings generating by the individuals who recorded the stimuli. We tested for relationships between the behavioral measures of interest (Empathic Accuracy, AQ score, and IRI-Empathic Concern score) and individual subject classification accuracies in each of the ROIs by computing Pearson correlations (testing for positive relationship between neural classification accuracy and behavioral measures of EA and IRI-EC and negative relationships between classification accuracy and AQ score). No reliable relationships were observed.
4.1 INTRODUCTION

In the preceding chapters, I report several lines of research investigating theory-like inferences about others’ emotional states. These experiments jointly suggest that abstract properties of eliciting events not only explain behavioral expectations and attributions of emotions throughout the lifespan, but also structure neural representations of emotions in regions involved in mental state reasoning. In this final section, I first summarize the main conclusions of each chapter, and expand upon some open questions raised by each. I then take a step back to evaluate the relationship between the developmental and neural methods employed in this thesis, and consider ways in which studies of cognitive neuroscience and cognitive development can be mutually informative. Using the experiments from this thesis as examples, I discuss how an interdisciplinary effort involving neuroscience, social psychology and developmental science could provide a more integrated picture of the cognitive mechanisms underlying emotion perception and inference. I end by discussing limitations and challenges of this overarching approach.

4.2 SUMMARY AND DISCUSSION

4.2.1 Chapter 1 Summary

In Chapter 1, I argued that the medial prefrontal cortex (MPFC) contains an abstract neural code that generalizes across emotions perceived in overt expressions and emotions inferred from context. Specifically,
neural patterns distinguished emotional valence (positive vs. negative trials) for both facial expressions and for short animations depicting emotion-eliciting events, and a classifier trained on neural patterns for one stimulus type could reliably classify the valence of the remaining type. Reliable cross-stimulus decoding indicates that there is valence information in MPFC that abstracts away from stimulus-specific features, whereas in other regions, such as the fusiform gyrus, neural representations of emotion appear to be specific to overt facial expressions. Interestingly, a region of the posterior temporal cortex seems to contain an intermediate representation that is invariant to perceptual modality (Peelen et al., 2010), but does not generalize to emotional valence when it is inferred from situational information. Together, these results suggest that perceiving and inferring others’ emotions evokes valence-specific patterns of neural activation at varying levels of abstraction throughout face-selective and theory of mind brain regions.

4.2.2 Limitations & Open Questions

Domain-Specificity: A question left open by this research concerns the precise relationship between abstract valence representations observed in MPFC (see also Bestelmeyer et al., 2014; Chikazoe et al., 2014; Kim et al., 2015; Peelen et al., 2010) and other hypothesized functions of the region. A challenge for understanding the functional organization of MPFC is that it is a large and potentially heterogeneous swath of cortex (see Bzdok et al., 2013), and even regions that are functionally specific within an individual may exhibit spatial heterogeneity across individuals (see Rebecca Saxe, Brett, & Kanwisher, 2006). The second experiment in Chapter 1 did reveal tentative differences along the dorsal-ventral axis of MPFC: there was evidence in the middle, but not dorsal subregion that the valence code generalized across experienced and attributed emotional states (i.e. the subject winning or losing money vs. observing a character experience a positive or negative outcome). However, dorsal

---

11 Many studies implicate MPFC in representing social or emotional information (Amodio & Frith, 2006; Chavez & Heatherton, 2015; Etkin, Egner, & Kalisch, 2011; Harris, McClure, Bos, Cohen, & Fiske, 2007; Hynes, Baird, & Grafton, 2006), but regions of MPFC are also recruited for a range of seemingly dissimilar tasks (Buckner & Carroll, 2007; Gross et al., 2014; Kampe, Frith, & Frith, 2003; McNamee, Rangel, & O’Doherty, 2013; Rodriguez, 2010; Rebecca Saxe, Moran, Scholz, & Gabrieli, 2006; Spreng, Mar, & Kim, 2008). It is unclear whether these various tasks activate the same portion of MPFC that is involved in emotion processing, and whether different subregions of MPFC all play the same role in third-person emotion attribution.
and middle subdivisions did not reliably differ from each other (i.e. cross-task discrimination accuracy was not reliably higher in MMPFC), making it difficult to interpret strong functional divisions between these regions.

A related challenge is that individual ROIs were selected from three searchspaces (ventral-middle-dorsal) that divide the MPFC rather arbitrarily, and could miss relevant functional distinctions (see Bzdok et al., 2013). For example, many have proposed a posterior-to-anterior abstraction gradient in prefrontal cortex (see Badre, 2008), and there is some evidence that this abstraction pattern extends to the representation of reward value/valence in medial regions (see Clithero & Rangel, 2013). For example, McNamee and colleagues (2013) found that whereas more posterior regions of VMPFC contained distributed neural patterns discriminating the values of specific stimulus categories, an anterior subregion seemed to contain a more general reward representation that generalized across stimulus category. While the research in Chapter 1 was among the first to directly assess the role of domain-general reward processes in emotion attribution, and finds some evidence for both shared and attribution-specific representations, the precise relationship between these mechanisms remains to be understood. The results are certainly consistent with the well-documented role for MPFC in mental state processing, but remains unclear whether the computations performed in each of the subregions are specific to the social domain, and further research is needed to characterize the interaction between domain-specific and domain-general processes in the identification of others’ emotional states.

**Causal Role:** A second challenge for future research would be understanding the precise causal role for MPFC in emotion attribution behaviors. Given the correlational nature of fMRI analyses, it is difficult to distinguish the representation of a given distinction (e.g. valence) from differences in processing that result from that distinction. For example, based on the decoding results from Experiment 1 alone, it is possible that subjects engage in different emotion regulation strategies when viewing the positive and negative stimuli, and that patterns of activity that distinguish valence actually reflect downstream differences in regulatory processing (see Buhle, Silvers, et al., 2013; Etkin, Egner, & Kalisch, 2011). Because the representation in dorsal MPFC does not generalize across third-person and first-person experiences, it seems unlikely that the response in this region reflects a fully general representation related to the subject’s own affective state. Still, there could be downstream
consequences of the emotion representation that are attribution-specific (e.g. feeling sympathy), but distinct from the representation of the emotional state itself. While lesion research confirms a causal role for MPFC in some aspect of understanding or responding to the emotional states of others (see review in Hillis, 2014), it is unclear from these data whether MPFC plays a role in the initial attribution, or in downstream motivational or empathic processes.

**Information Flow:** Finally, one of the most exciting directions for future research will be characterizing information flow between the different brain regions that exhibit valence-specific neural patterns: how is an abstract representation of another’s emotional state constructed from diverse sensory inputs? An obvious hypothesis is that the levels of abstraction reported in Chapter 1 constitute a processing hierarchy, where specific inputs (e.g. facial expressions) are initially processed in modality-specific perceptual regions (e.g. fusiform face area, superior temporal sulcus), followed by integration in multimodal regions (e.g. posterior superior temporal cortex). The conceptual valence information decoded in MPFC could then reflect the final, most abstract stage of representation within this cortical hierarchy. It is also possible that higher-level representations engage in feedback that modulates the response in modality-specific regions, perhaps enhancing attention or altering perceptual processes strategies.

Can fMRI be used to characterize how different regions interact in the process of emotion attribution? Testing hypotheses about information flow between cortical regions will require creative methods that overcome the temporal limitations of fMRI. Excitingly, Anzellotti and Saxe recently developed a new set of methodologies for characterizing interactions, and directionality of interactions, between brain regions, and applied them to the data from Chapter 1. They find that top-down modulations from regions like MPFC do influence the visual processing of faces in inferotemporal cortex (Anzellotti and Saxe, personal communication). In principle, these methods could also be used to characterize how brain regions associated with different sensory modalities (e.g. auditory and visual emotional stimuli) interact to yield the multimodal and conceptual representation in higher-level regions.

**4.2.3 Chapter 2 Summary**
Whereas the research begun in Chapter 1 aims to understand how abstract emotion representations emerge over stages of cortical processing, Chapter 2 aimed to explore how abstract, conceptual building blocks for emotion inference emerge over human development. In particular, given that some emotion sensitivity is evident in infancy, what is the content and scope of early emotion representations? In Chapter 2, I described a new line of research probing the relationship between infants’ knowledge of emotions and goal-directed actions. Despite a rich literature on early perceptual discrimination of emotional expressions, it was unknown whether infants recognize that these emotional reactions are caused by events or situations in an agent’s environment.

Across several studies, looking time measures were used to assess 8 and 10-month-old infants’ expectations about how an agent should respond emotionally to attaining or failing to attain a demonstrated goal. At 10 months, there was reliable evidence that infants form expectations about how an agent would react to goal completion, as looking time to the same emotional expression varied depending on whether the expression was congruent with the preceding action outcome. This phenomenon generalized across two distinct goal contexts, and a control experiment suggested that these expectations depended on an analysis of the agent’s action as directed towards a stable goal. At 8 months, there was some evidence for this ability, but the findings were less robust in that they did not generalize to the second goal context. Interestingly, the extent of the congruency effect seemed to differ between completed and failed goal outcomes; infants looked longest when a negative emotional reaction followed a completed goal, whereas a positive emotion following a failed goal did not elicit a strong violation of expectation response. This asymmetry could reflect either a limit to infants’ knowledge of the relation between goal outcomes and emotions, or a set of rational assumptions about the goal context (e.g. that a failed attempt can simply be repeated).

Together, these results suggest that infants have some knowledge of the conditions that elicit different emotions, and in particular of how emotional states relate to outcomes of goal-directed actions. The data serve as an initial step towards understanding the scope of infant emotion knowledge, and provide the first evidence that the representations involved in action processing in infancy (Csibra & Gergely, 2007; Woodward, 1998) are sufficiently abstract to support expectations about emotional reactions. Evidence for emotion understanding in
preverbal infants highlights the importance of emotions for learning about and from other people (Campos, 1983), and suggests that emotional expressions may provide a rich source of information throughout the lifespan.

4.2.4 Limitations and Open Questions

Granularity and breadth: While these findings show that infants perceive emotional reactions in relation to preceding events, there is much that remains unclear about the sophistication and scope of this knowledge. For example, Chapter 2 focused on only a single binary distinction between positive and negative valence, rather than on more fine-grained emotional distinctions (e.g. discriminating events that would elicit anger vs. sadness). Moreover, these experiments focused on a single class of events that could elicit positive or negative affect (failure or success of a simple instrumental goal). Do infants also represent the affective consequences of other social, physical, or goal-related outcomes? Given that infants’ sensitivity to emotional congruency was somewhat limited even in the context of simple goals, it seems plausible that emotion knowledge in the first year is rather narrow. However, it is also possible that infants’ would exhibit richer or more complete knowledge of the causes of emotional expressions in other contexts (e.g. as used for social affiliation or communication).

Inferential role: Future research could also more fully characterize the conceptual role of emotion concepts at 10 months and beyond. When and in what way are emotional expressions integrated into other aspects of social cognition? For example, the present research asked whether infants reason about the events that precede different emotional reactions. Do infants also have knowledge of the consequences of different emotions such that they can predict subsequent actions or behaviors? A few studies have begun to tackle this question (e.g. studies on infants’ use of emotions to predict object-oriented actions: (Egyed, Király, & Gergely, 2013; Hepach & Westermann, 2013; Phillips, Wellman, & Spelke, 2002; Vaish & Woodward, 2010; Feiman et al., 2013), but as with the present paradigm, existing evidence is somewhat mixed. However, there are a number of experiments in which emotional expressions have been used to set up social relationships or establish shared preferences (e.g. Hamlin et al., 2013; Johnson et al., 2010; Mahajan & Wynn, 2012). For example, by 9 months of age, infants can use emotional vocalizations towards different food items to infer preferences and establish self-similarity, and use
the shared or divergent preferences to guide subsequent social evaluations (Hamlin et al., 2013). Thus, although not directly investigating the development of emotion understanding, these studies provide hints that infants use emotional displays to make a range of social inferences.

**Flexibility:** A related question is whether infants have a flexible representation of the causal structure of expressions and their eliciting situations, or whether they might detect some simpler incongruence between the valence of the event and the valence of the subsequent emotional display. On the latter interpretation, infants would form some abstract representation that encodes what is in common across a successful outcome and a positive expression, but this knowledge could fall short of a mature, theory-like understanding of the causal relation between goal outcomes and affective responses. One way to distinguish these possibilities would be to investigate whether infants can flexibly integrate other relevant information when attributing emotions to others. For example, can preverbal infants take into account an agent’s perceptual access and what the agent knows or believes has occurred when reasoning about how the agent will react to an event? (see Repacholi, Meltzoff, & Olsen, 2008 for evidence of this ability at 18 months).

**Relation to prosocial behavior:** Finally, knowledge of emotions and emotion-eliciting situations could serve as essential conceptual building blocks supporting our species’ unique prosocial and cooperative tendencies (see Brownell, Svetlova, Anderson, Nichols, & Drummond, 2013; Thompson & Newton, 2013; Vaish & Warneken, 2011). A large number of studies report that humans have preferences and expectations about prosocial events from early in infancy (Geraci & Surian, 2011; Hamlin, 2014, 2015; Hamlin, 2013; Hamlin et al., 2013; Hamlin, Wynn, Bloom, & Mahajan, 2011), and that as soon as they are physically capable, children begin engaging in actions for the benefit of others (Brownell, Svetlova, & Nichols, 2009; Liszkowski, Carpenter, Striano, & Tomasello, 2006; Warneken & Tomasello, 2006, 2007, 2009). A tempting interpretation of these behaviors is that early helpful or prosocial tendencies depend on social-cognitive representations of the kind examined in this thesis. Do toddlers, in fact, reason about the affective consequences of various outcomes when engaging in goal-congruent or “prosocial” actions for others (Davidov, Zahn-Waxler, Roth-Hanania, & Knafo, 2013; Dunfield, 2014; Dunfield & Kuhlmeier, 2013), or are these behaviors driven by a simpler set of attentional or motivational
factors (see Kenward & Gredebäck, 2013; Paulus, Kühn-Popp, Licata, Sodian, & Meinhardt, 2013)? Investigating the role of early emotion concepts in empathic and prosocial behavior could be helpful for understanding the extent to which these abilities rely on mechanisms that are continuous throughout the lifespan.

4.2.5 Chapter 3 Summary

Chapter 3 aimed to elucidate the specific properties of events that support inferences about others’ emotional states, and to examine how such event properties features might structure the representation of attributed emotions in the human brain. In other areas of cognitive neuroscience (e.g. the study of visual object recognition: Carlson et al., 2013; Carlson, Tovar, Alink, & Kriegeskorte, 2013; Charest, Kievit, Schmitz, Deca, & Kriegeskorte, 2014; Cichy et al., 2014; Kriegeskorte, Mur, Ruff, et al., 2008; Liu et al., 2013; Peelen, He, Han, Caramazza, & Bi, 2014; D. Yamins, 2014), researchers have been able to characterize the feature or dimensions that organize neural representations at different levels of a cortical hierarchy (e.g. representations supported by different population codes throughout inferotemporal cortex). A goal of this thesis is to show that these approaches (see reviews in Haxby, Connolly, & Guntupalli, 2014; Kriegeskorte & Kievit, 2013) can be fruitfully applied to domains of abstract, high-level cognition. The findings from Chapter 3 demonstrate that it is indeed possible to make progress in characterizing the representational organization of brain regions involved in emotion attribution.

Based on the combination of neural and behavioral results, I suggest that our knowledge of others’ emotions structured around a space of abstract, high-dimensional event features, that brain regions selective for mental state reasoning support relatively subtle emotional distinctions, and that the neural representations in these regions cannot be explained in terms of primitive affective dimensions or categories. Specifically, a set of 38 abstract features captured properties of events that vary between different emotions, and these features contained sufficient information to approach human behavioral performance when classifying the emotion labels of different stimuli. This set of features constitutes one of the first attempts to explicitly model and quantitatively evaluate the content of our intuitive theory of emotion. In the neural data, Representational Similarity Analyses showed that this high dimensional feature space was best able to capture representations in theory of mind brain regions, and
neural patterns in these regions were not reducible to more primitive affective dimensions such as valence and arousal.

4.2.6 Limitations and Open Questions

**Feature space & algorithm:** A key limitation of this research is that the set of 38 event features was neither discovered from the data, nor optimized to best capture variance across the twenty emotion categories. Based on the results of Chapter 3, I suggest that a space of this kind (i.e. abstract properties of emotion eliciting events) will be a fruitful direction for characterizing the structure of human emotion knowledge, but these results by no means demonstrate that the set of features use in the present research is optimal or adequate. These features were motivated by prior literature on appraisal theory (Fontaine, Scherer, & Soriano, 2013; Meuleman & Scherer, 2013; Scherer & Meuleman, 2013), but were nonetheless compiled in a somewhat haphazard, intuition-driven manner. There are likely features important for certain discriminations that were missed altogether in the present feature set and, as suggested by the dimensionality reduction analysis, features that are potentially redundant or uninformative. Although this feature space outperformed competing models in capturing the distinctions between different emotions, an important challenge for future research would be both expanding and pruning this space. Importantly, the general framework used in Chapter 3 lends itself well to iterating on the candidate feature spaces and comparing the performance of newly proposed basis sets.

In addition to optimizing the features themselves, there may also be room to improve the algorithm that was used to classify the twenty emotion categories from these feature spaces. I used a relatively simple classification model (a linear SVM) that treats the emotion class as a linear combination of feature weights, and it is possible that this classifier underfit the data in the current context. It is very plausible that there are nonlinearities in the relation between event properties and the emotions we ascribe. With a more flexible model (e.g. an algorithm that could learn complex interactions between features), the existing feature space might support discriminations that more closely approach human performance (Meuleman & Scherer, 2013).

**Representational Format:** Perhaps more important than either the feature space or the learning algorithm is the representation used to encode the emotion-specific knowledge itself. In Chapter 3, each
candidate feature space used a flat, unstructured representation: each emotion was encoded as a simple vector of 
the ratings for each feature, averaged across exemplars. This representation lacks compositionality and cannot 
naturally encode relational structure, meaning that ignores the temporal and causal relations between different 
aspects of the event. A less impoverished representation, particularly one that could encode a logical, causal 
theory, seems essential for adequately explaining human knowledge in this domain. For example, consider a 
situation similar to the event described above: Joan is embarrassed because she spilt a red drink on her white dress 
while at an upscale work event. Attributing embarrassment to Joan involves not only recognizing that there were 
other people present and that the event was inconsistent with Joan’s goals (the kinds of features encoded in the 
current 38-dimensional space), but also understanding that those people stand in the appropriate epistemic 
relation to the event (i.e. that other people saw the dress covered in wine before Joan could gracefully exist), and 
hold a particular social relation to Joan (i.e. a professional hierarchy, potentially including Joan’s superiors).

The work reported in this thesis is strongly motivated by the theory-theory of concepts—the notion that 
conceptual representations must be understood in terms of their computational role within a coherent 
explanatory theory (Carey, 1985, 2009; Gopnik & Wellman, 1994; Keil, 1992; Morton, 1980). Yet, the actual 
knowledge representation used in Chapter 3 is far more analogous to the way concepts have been defined in 
prototype theory (i.e. graded categorization based on feature similarity to some category prototype or centroid; 
Rosch, 1973; Shaver, Schwartz, Kirson, & O’Connor, 1987). By relying on a simple vector of feature weights (e.g. 
Did the event involve the target directly? Were other people present for the event?), the current approach fails to 
capture important nuances in the structure of our intuitive emotion knowledge. In other domains of cognitive 
science, including social cognition, researchers have shown that it is possible to model rich intuitive theories using 
formalisms that capture the structured, hierarchical nature of human conceptual knowledge (Baker et al., 2009; 
Tenenbaum et al., 2011). Thus, the current feature space could simply serve as inspiration for future approaches 
that bring more powerful representational systems to bear on modeling and explaining human emotion inference.

**Stimulus-computability:** A final shortcoming of the present approach is that although it allowed for 
quantitative comparison between different emotion theories (i.e. circumplex model, basic emotions model,
appraisal model), none of the candidate “models” were stimulus-computable (meaning they could not be applied directly to the text stimuli itself). Rather, human subjects were required as an intermediate step in the process, rating each stimulus on the features corresponding to each model. With a sufficiently complete understanding of event semantics, natural language, and social cognition, this step could, in principle, be performed without a human in the loop. In other words, rather than a subject rating features such as “to what extent was this event caused by an agent?”, these event properties could be inferred algorithmically from the text itself. While such advanced natural language processing capacities are likely a long way off, recent interest in automated emotion recognition in the machine learning and artificial intelligence community has led to the exploration of novel natural language approaches to emotion classification. For example, Socher and colleagues (2013) recently used a Recursive Neural Tensor Network trained on emotion labeled sentence parse trees (which, in contrast to simple word frequency vectors, represent the structure of phrases within the sentence) to break previous classification benchmarks on classic tasks in sentiment analysis (Pang & Lee, 2004)12. Exciting questions for future work include whether this approach can extend to more fine-grained emotional distinctions beyond valence (see Poria et al., 2015; Smith & Lee, 2013), and whether such an approach would benefit from an intermediate stage that encodes the text input in terms of abstract event features of the kind explored in this thesis (as opposed to training on emotion labels directly; see Agarwal & Mittal, 2013; Cambria & Howard, 2014 for interesting directions). In fact, it is possible that even in existing models (e.g. Socher and colleagues’ RNN), certain the intermediate layers could resemble the kinds of properties or dimensions explored in Chapter 3.

4.3 MOTIVATING THE APPROACH

In chapters 1-3, I present data that bear on several distinct questions regarding the neural basis of emotion inference, and its origins in human development. Because these studies draw on a number of literatures and methodologies—specifically from neuroscience, affective science, and developmental psychology—I

---

12 Interesting and impressive demo available here: http://nlp.stanford.edu/sentiment/
conclude by motivating this interdisciplinary approach more broadly. Thus far, I have interpreted the findings from each chapter in relative isolation, evaluating them with respect to the local hypotheses posed in each line of research. Here, however, I propose that neural and developmental data can also inform and constrain one another. First, I discuss how the complementary tools of neuroscience and cognitive development can, in principle, provide an integrated picture of the nature and origins of human cognitive abilities. I then discuss relationships between the lines of research reported in this thesis, and suggest ways in which the methods employed in the previous three chapters could provide convergent and complementary insights into the cognitive mechanisms supporting emotion inference and attribution.

4.3.1 Integrating neuroscience and developmental science

I argue that combining the theories and methods of neuroscience and cognitive development could shed light on four classes of outstanding questions in cognitive science. First, there is a long-standing debate regarding human cognitive architecture (Fodor, 1983; Hirschfeld & Gelman, 1994; Kanwisher, 2010): does the mind consist of a small number of general-purpose mechanisms carrying out diverse functions, or a larger number of functionally-specialized mechanisms tailored to the computational demands of specific domains? For any given cognitive phenomenon, we must characterize the contributions of domain-specific and general-purpose systems. Second, we must understand how the assumptions or principles of the domain are represented, and how the structure of those representations supports domain-relevant operations. I argue that to fully understand these systems, we must characterize the form and content of mature cognitive representations as well as the developmental processes that yield them. This brings us to a third question: what is the initial state of the human mind (i.e. what is specified genetically?) and what are the mechanisms that yield mature cognitive representations from these initial resources? Mechanisms of change could include both experience-independent maturational mechanisms, as well as learning mechanisms that yield qualitatively new representational systems (Carey, 1985, 2009). For any domain, we want to understand the interplay between innate structure and learning in the emergence of adult-like representations. A final, related question concerns possible continuities and
discontinuities in cognitive mechanisms over development. Does behavioral change correspond to the emergence of entirely new mechanisms, or on the tuning and refining of existing systems?

These four questions are among the most fundamental in cognitive science, and for each question, I argue that combining the vocabulary, tools, and phenomena of neuroscience and cognitive development can provide insights that go beyond what has been learned from either discipline alone. Most will agree that the ultimate goal of cognitive science includes an understanding of the human mind at multiple levels (e.g. computational, algorithmic, implementational; Marr, 1982). While Marr argued that different levels of analysis could be pursued independently, a premise of modern cognitive neuroscience is that these levels inform and constrain one another, and that the theories and phenomena of cognitive psychology can thus guide understanding of neural function (Frank & Badre, 2015). Although early research in cognitive neuroscience tended to draw on tasks and phenomena from adult psychophysics, research in other domains (e.g. theory of mind) has been strongly inspired by paradigms and conceptual distinctions from developmental psychology (see Brass, Schmitt, Spengler, & Gergely, 2007; Jastorff, Clavagnier, Gergely, & Orban, 2011; Ramsey & Hamilton, 2010; Saxe, Carey, & Kanwisher, 2004). Moreover, findings from developmental psychology generate hypotheses about the innate cognitive architecture, which can guide our understanding of functional organization in the brain, and provide signatures for identifying and dissociating different neural systems (see Dilks, Julian, Kubilius, Spelke, & Kanwisher, 2011; Hyde & Spelke, 2011; Platt & Spelke, 2009; Spelke & Lee, 2012).

Here, I suggest that this relationship between neuroscience and cognitive development is partially symmetric, and that the study of human development can also benefit from the theories and findings of neuroscience. For example, neuroscience methods could provide evidence that converges with results of behavioral studies (e.g. looking time), possibly ruling out alternative explanations of a behavior, or it could enable comparison with relevant populations (e.g. different species, age, developmental history), allowing us to “look under the hood” in populations otherwise difficult to study effectively. Perhaps most excitingly, neuroscience could provide evidence that differs qualitatively from what we learn from behavior. In particular, neuroscience can provide a unique window into intermediate stages in a perceptual hierarchy or processing stream (Freiwald 


allowing access to representations that may be encapsulated from awareness or even behavior. This is itself a deep contribution to cognitive science, as it provides constraints on the algorithms by which information is transformed during online processing and inference. Moreover, these neural findings can provide hypotheses and tools for charting the emergence of different levels of representation over human development, and can yield novel predictions for behavioral studies.

### 4.3.2 Emotion inference as a case study

Can we use findings from neuroscience to uncover the representations that support human emotion attribution, and employ the resulting knowledge to characterize the developmental origins of this ability? Because the research reported in this thesis is among the first to tackle causal emotion attribution, the current data, of course, fall short of this aim. Nonetheless, there are a number of ways in which the methods and conclusions of this kind of research could be combined to yield a more complete picture of the mechanisms underlying emotion perception and attribution. Here I discuss a few such relationships, linking the reported findings back to the fundamental developmental questions posed above.

#### 4.3.2a Domain Specificity

To explain and predict the behavior of other people, we rely on an intuitive theory that specifies abstract causal relations between different mental states (e.g. how desires, expectations, and beliefs determine emotions), between mental states and events or circumstances in the world (e.g. how a person’s perceptual experiences shape her beliefs), and between mental states and overt behavior (e.g. that people act so as to bring about their desires and intentions efficiently). On its face, this intuitive theory has properties in common with many domains of abstract knowledge. For example, we might rely on an intuitive theory of the weather to predict that it will rain (e.g. based on the cloudy sky), and to make inferences about where the storm will move (e.g. based on the wind).

A central question for the cognitive science of social cognition has been clarifying the extent to which there are cognitive mechanisms specific to processing social information (Apperly, Samson, & Humphreys, 2005; Gopnik & Wellman, 1994; Hirschfeld & Gelman, 1994; Saxe, Schulz, & Jiang, 2006; Scholl & Leslie, 1999; Stone & Gerrans, 2006), and characterizing the precise domain and scope of such mechanisms if they exist (Apperly &
Butterfill, 2009; Blair, 2005; Bruneau et al., 2012a; Burnett, Bird, Moll, Frith, & Blakemore, 2009; Doré, Zerubavel, & Ochsner, 2015; Lieberman, 2007; Mitchell, 2009; Saxe & Powell, 2006; Saxe, Schulz, et al., 2006). Is the brain equipped with distinct selective mechanisms for different domains of abstract knowledge (including the domain of other minds), or are diverse domains of reasoning supported by general-purpose mechanisms that support prediction, inference, and explanation regardless of content?

While a domain-general mechanism might seem more plausible, a number of lines of evidence suggest that reasoning about other people is “special”. Inferences about other people’s mental states exhibit a unique and systematic developmental trajectory (Wellman, Cross, & Watson, 2001), are specifically impaired in individuals with Autism Spectrum Disorders (Simon Baron-Cohen, 2000), and activate a selective network of cortical regions (the “Theory of Mind” network: MPFC, PC, bilateral TPJ, and bilateral ATL: Dodell-Feder, Koster-Hale, Bedny, & Saxe, 2011; Saxe & Kanwisher, 2003). The neural evidence for a domain-specific mechanism is particularly strong: these theory of mind (ToM) regions can be identified in 80-90% of individual subjects, and the selective response is robust and replicable across a wide range of stimuli and tasks (e.g. Gallagher et al., 2000; Sommer et al., 2007; see Koster-Hale & Saxe, 2013 for review). In the RTPJ, the neural response is most selective, responding maximally to stimuli that elicit reasoning about representational mental states such as beliefs (Dodell-Feder et al., 2011; Downing, Jiang, Shuman, & Kanwisher, 2001; Pelphrey & Morris, 2006; Saxe & Powell, 2006; Saxe, Schulz, et al., 2006; Saxe, Xiao, Kovacs, Perrett, & Kanwisher, 2004).

While there is reliable evidence that this set of regions is recruited during mental state reasoning, the bounds and scope of this network have remained controversial. If there is universal, selective neural substrate for the social cognition, what exactly are the representations and computations that fall within the scope of that mechanism? Answering this question requires going beyond testing which classes of stimuli drive the overall recruitment of the network, and instead probing the precise features or dimensions that are represented by neural populations within it.

One hypothesis in the literature has been that the ToM network is involved in representing mental states such as beliefs or second-order intentions, whereas a set of lower level systems support a more “direct”
representation of mental states like emotions. For example, on one popular proposal, we read others’ emotions by “simulating” them using neural mechanisms generally engaged in the first-person experience and expression of emotions (Gallese & Sinigaglia, 2011). On this view, we understand emotions via complementary activity from the canonical “mirror neuron system” (MNS: IPS, IFG, SMA, somatosensory cortex), a set of regions argued to be involved in bodily simulation of emotion-specific facial movements, and a putative secondary mirroring system, a set of limbic and paralimbic regions (e.g. insula, amygdala, cingulate gyrus) argued to be involved in simulating the underlying affective states (Gallese, 2007; Keysers & Gazzola, 2014; Lawrence et al., 2006; Molenberghs, Cunnington, & Mattingley, 2012; Pfeifer et al., 2008; Sel, Forster, & Calvo-Merino, 2014; van der Gaag et al., 2007). By these accounts, the ToM network implements an alternative, more sophisticated system that may be useful in specific, relatively infrequent contexts (e.g. explicitly explaining a person’s behavior in terms of his or her false belief), but that day-to-day social cognition involves a set of simpler representations (of emotions, actions) that are supported by lower-level simulative processes (Gallese & Sinigaglia, 2011). Another proposal has been that cognitive and emotional empathy (reasoning about beliefs and emotions, respectively) depend on two distinct cognitive mechanisms, implemented in separate neural circuits and exhibiting double dissociations across psychopathologies (Blair, 2005).

Thus, according to these various theories, the domain of the ToM network is limited to a relatively small corner of epistemic mental states. If correct, this account would provide an important constraint on scope of computations supported by this domain-specific ToM mechanism. Such a division in the cognitive architecture supporting different aspects of social cognition would be of interest to researchers investigating social cognition across disciplines. However, despite many claims that affective and epistemic perspective-taking are supported by distinct neural systems (Blair, 2005; Shamay-Tsoory et al., 2004; 2009; van der Gaag et al., 2007), the tasks used to study emotion and belief reasoning have differed both in the specific domain (epistemic states like beliefs vs. emotional states), in the inferential demands of the task (inferring states from abstract scenarios or verbal vignettes vs. labeling/categorizing overt expressions), and possibly in the depth of mental state processing that was elicited (e.g. perceptually discriminating or labeling vs. explicitly considering how the target feels).
To understand the relationship between emotion attribution and other aspects of mental state reasoning, and to clarify the domain of computation supported by the ToM regions described above, we must manipulate the domain of social reasoning without varying the inferential demands. We can do so by studying emotion attribution in a context that requires theory-like inferences based on abstract situational information. Indeed, when we use a rich space of emotions that must be inferred from abstract situation information (Chapter 3), we find remarkable overlap in the regions that support emotion classification and those traditionally associated with theory of mind.

Previous claims of neural dissociation between affective and epistemic theory of mind may have been the consequence of operationalizing “affective” theory of mind in a way underestimates the richly inferential nature of this ability. A goal of this thesis is to highlight the fact that emotion knowledge is much more than perception of overt displays; it involves a rich, causal theory that integrates a variety of social-cognitive representations to support inferences in absence of any perceptual inputs. When we design experiments that tap this aspect of emotion knowledge, we find that the ToM network is strongly and selectively implicated. This result suggests that emotion attribution is indeed within the scope of the ToM network (and this holds for stimuli that range from facial expressions, to animations, to verbal vignettes), consistent with the view that this set of regions supports a domain-specific system dedicated to theory-like reasoning, prediction, and explanation of others’ mental states. Thus, studies probing the fine-grained neural representations within different regions can be used make progress on important questions regarding domain specificity in human cognition.

4.3.2b Representations and their development

In order to explain how abstract concepts are constructed, we must characterize the representations and computations that underlie mature reasoning in adults, in order to posit mechanisms that bridge from some initial state to that mature state (Carey, 2009). The research described in Chapter 3 provides a first pass characterization

---

13 Of course, the theory of mind network is made up of several regions that presumably differ in their precise computational roles. Future research might aim to understand how different components of the ToM network differentially contribute to the process of emotion perception and inference.
of the content and dimensionality of adults’ intuitive knowledge of emotion. Both the behavioral and neural results suggest that space of emotions people attribute is best characterized by abstract properties of different emotions’ antecedent causes. Moreover, this work provides an initial sketch of a particular set of event properties that are reliably diagnostic of different emotional states.

Could this work provide a useful framework for investigating the development of human emotion knowledge? For example, although the simpler affective spaces like the circumplex model or the basic emotions model failed to capture the range of emotions that are attributed by human adults, it is possible that a simpler basis may be sufficient to capture the space of emotions that are represented earlier in development. Indeed, basic emotion theory was highly motivated by evolutionary, developmental and cross-cultural perspectives (Ekman et al., 1987; Ekman & Friesen, 1971; Izard, 1992; Shariff & Tracy, 2011), and might be a valid account of the representational resources available innately. The results of Chapter 3 suggest that the six basic emotions do not provide an adequate set of conceptual primitives for mature emotion inference, in that they do not provide the building blocks necessary to explain the full range of mature emotion concepts. These basic emotions might nonetheless be important developmental primitives that are combined with other social-cognitive representations in the construction of more complex theories of emotion.

The approach used in Chapter 3 could be useful for characterizing the emotion representations available innately, but also for understanding the emergence of new emotion knowledge over development. A large body of research has studied when different emotion concepts emerge in childhood, documenting stark differences in the developmental trajectories for different emotions (Banerjee, 2002; Ferguson, Stegge, & Damhuis, 1991; Nunner-Winkler & Sodian, 1988; Russell, 1990; Saarni, 1999; Wellman & Banerjee, 1991; Widen & Russell, 2010; Yuill, 1984). However, one limitation of much of this research is that it is not clear whether children fail to attribute an appropriate emotion to a situation because they lack the social-cognitive or semantic knowledge necessary to parse the relevant properties of the event, or because they do not understand the emotion that those event properties elicit, or both.
The 38-dimensional feature space constructed in Chapter 3 provides a basis for teasing apart these alternatives: when children attribute inappropriate emotions to a situation, do they do so because they fail to ascribe the relevant properties to the event, or because they lack the theoretical knowledge that maps between abstract event properties and different emotional states? For example, it is not until age 8 that children reliably recognize that a person will feel embarrassed in prototypical scenarios (e.g. “Joan spilled grape juice all over her white dress. All the kids laughed at her”; Widen & Russell, 2010). Using its appraisal profile from Chapter 3, we can articulate the event regularities that are diagnostic of an embarrassed state. For example, compared to events that elicit devastation or disappointment, subjects rated the embarrassing events as more likely to be caused by an agent, more likely to be within control of the target character, more likely to involve the target character directly, and more likely to involve other people. With an explicit model of the event properties that are diagnostic of particular emotions (according to adult subjects’ intuitive theories), we could then ask whether developmental differences in the attribution of embarrassment reflect change in the identification of these properties (e.g. understanding that spilling the juice was Joan’s fault and she might have avoided it by being more careful; recognizing that a purple stain on a white dress will be visually salient and draw the attention of people around Joan), or change in children’s understanding of the emotion that follows from these more abstract conditions. In previous studies on children’s recognition of different facial expressions, researchers have made progress in understanding children’s perceptual confusions (e.g. fear with surprise, disgust with anger) by analyzing the similarity in facial action units between different pairs of expressions (Camras, 1980; Wiggers, 1982). Having an explicit account of the eliciting conditions that are similar and dissimilar between different emotions could provide a similar framework for understanding the errors children make in attributing emotions based on situational information. In general, by decomposing our knowledge of emotions into a set of simpler (but still abstract) regularities in the eliciting conditions, we can make headway in understanding which aspects of this
intuitive knowledge of emotions are available early in development, and what mechanisms of change might be necessary for expanding the scope of this early system into a full-fledged theory of others’ emotions14.

4.3.2c Innateness and continuity

Finally, the research described in this thesis provides some preliminary tools for investigating whether emotion attribution abilities depend on a developmentally continuous mechanism. The designs used in Chapters 1 and 2 are actually quite comparable, as both involve reasoning about the valence of simple animated events in which an agent completes or fails to complete a goal. One obvious question is whether the MPFC mechanism documented in Chapter 1, and further characterized in Chapter 3, supports the behaviors observed in Chapter 2.

As discussed above, it remains unclear exactly what infants represent about emotion eliciting situations, or whether they have a fully integrated understanding of different emotions’ antecedent causes. Still, to detect the congruence between the events, they must have some representation of valence that captures what is in common across a positive or negative expression and a positive or negative goal outcome. In Chapter 1, MPFC was the only region observed to contain such a representation and, at least in adults, this region appears to support a very abstract, theory-like space of emotion knowledge (Chapter 3). Unfortunately, multivariate methods alleviate only some of the spatial constraints of univariate methods—MVPA still tends to rely on relatively low-frequency information (Freeman, Brouwer, Heeger, & Merriam, 2011; Op de Beeck, 2010), meaning that classification can be seen as a lower-bound on the information available in a given region (Kriegeskorte & Kievit, 2013). Thus, the present methods could be insensitive to emotion information in other regions (e.g. the amygdala) if that

14 Although Representational Similarity Analysis (the method used in Chapter 3) has been popularized in the context of neuroscience research, many of the virtues of the approach could extend to developmental studies as well. A primary advantage of RSA is the ability put various different models/theories onto a fair (parameter-free) playing field by caching out each into a matrix representing the pairwise dissimilarities between every pair of stimuli or conditions. These can then be correlated with some target matrix—usually derived from similarity within some neural population—to determine what model or feature set best explains that target space. In fMRI research, that similarity space is usually computed from voxel patterns, but it could also come from neurophysiological data, from a behavioral confusion matrix, or from any other data source that allows for computing similarity. By moving from raw data space into stimulus/condition space, RSA eliminates the need for establishing correspondence in the input; in neuroscience, this has been useful because it allows us to compare across individuals and species without putting brain data into correspondence. From a developmental perspective, it could similarly provide a simple framework for comparing different types of data; for example, one could obtain pairwise similarity data from explicit similarity judgments in adults, and pairwise novelty/discrimination data in infants, and compare their fit with various quantitative models.
information is represented at a smaller spatial scale (see Gothard, Battaglia, Erickson, Spitler, & Amaral, 2007; Hadj-Bouziane et al., 2012).

In fact, the default hypothesis, (e.g. see proposals in Farroni et al., 2007; Johnson, 2005) has been that that the emotion sensitivities observed in infancy reflect the operation of primitive, automatic mechanisms for detecting survival-relevant stimuli (Buhle, Kober, et al., 2013; Dolan & Vuilleumier, 2003; Happé & Frith, 2014; Zald, 2003). Given that the prefrontal cortex is known to undergo prolonged developmental change, it has been assumed that innate perceptual sensitivities to emotion must reflect lower-level attentional mechanisms supported by innate subcortical circuits. Further evidence that the MPFC mechanism might be developmentally discontinuous is that the response to emotion processing or mentalizing in MFPC exhibits change over childhood and adolescence (Blakemore, 2008; Burnett et al., 2009; Crone & Dahl, 2012; McRae et al., 2012; Monk et al., 2003; Moriguchi, Ohnishi, Mori, Matsuda, & Komaki, 2007; Somerville, Fani, & McClure-Tone, 2011; Yurgelun-Todd & Killgore, 2006), and there are robust developmental differences in MPFC connectivity to other emotion-related systems such as the amygdala (Gabard-Durnam et al., 2014; Gee et al., 2014).

Another possibility is that an infant’s capacity to identify others’ emotions is supported by distinct, early emerging emotion representations in temporal regions (Blasi et al., 2011; Grossmann, 2005; 2010a). For example, in adults, a region of the superior temporal sulcus is sensitive to the match between a target’s preference (conveyed via an emotional expression towards an object) and subsequent goal directed action, showing a heightened response to emotionally incongruent actions (Vander Wyk et al., 2009). One highly speculative interpretation of these data would be that the superior temporal sulcus contains relatively simple, but integrated representations actions, goal outcomes, and emotional expressions; in that case, infants’ sensitivity to the incongruent reactions in Chapter 2 could be supported by this kind of intermediate mechanism that is neither an evolutionarily ancient subcortical circuit, nor the full fledged adult-like system. On this hypothesis, the medial

15 Interestingly, in Chapter 1, I found that the STS contained representations of emotion in both facial expressions and situations, but that the population code did not generalize across these stimulus types. However, Vander Wyk and colleagues use a repetition suppression/prediction error design that is less limited by spatial scale constraints.
prefrontal cortex might come online later as children construct a more complete causal theory of the eliciting conditions for different emotions (see Happé & Frith, 2014).

Of course, a final possibility is that the emotion knowledge observed in infancy is supported by the same medial prefrontal mechanism that operates in adulthood. While there is change in MPFC activity over development, there is nonetheless reliable evidence that MPFC is recruited during mentalizing tasks as early as 4 years (Blakemore, 2008; Bowman, Liu, Meltzoff, & Wellman, 2012; Gweon, Dodell-Feder, Bedny, & Saxe, 2012; Kobayashi, Glover, & Temple, 2007; Sabbagh, Bowman, Evraire, & Ito, 2009; Saxe, Whitfield-Gabrieli, Scholz, & Pelphrey, 2009; Sommer et al., 2010). In fact, in some of these studies, children actually exhibited greater activity in the MPFC, relative to adults performing the same task. Moreover, the assumption that a functional MPFC emerges late in human development has come under scrutiny based on a number of recent findings; for example, based on a review of recent fNIRs results (Grossmann et al., 2008; Grossmann, Parise, & Friederici, 2010; Grossmann & Johnson, 2010; Lloyd-Fox, Blasi, Everdell, Elwell, & Johnson, 2011; Naoi et al., 2012; Tzourio-Mazoyer et al., 2002), Grossman has argued that MPFC activates to social or emotional stimuli in early infancy, and even in newborns (Grossmann, 2013b). In fact, there have been several reports of infants exhibiting ventromedial prefrontal activity in response to emotional expressions in particular (Blasi et al., 2011; Grossmann, 2010a; Minagawa-Kawai et al., 2009).

In sum, it is unknown whether infants’ ability to detect, recognize, and interpret others’ emotions is based on the same underlying mechanism that supports rich, emotion-related reasoning throughout adulthood. The recent findings on the role of MPFC in early social cognition (see reviews in Grossmann, 2013a, 2013b) raise exciting the exciting possibility that this cortical mechanism supports emotion-related perceptions and attributions continuously throughout the development. However, the research in Chapter 2 probed only very basic emotion concepts, and found that infants’ understanding even of the relation between goal outcomes and positive versus negative expressions may have been incomplete at the ages tested. Thus, it is remains possible that these early expectations about emotional reactions are supported by a simpler, more limited mechanism. Nonetheless, integrating the methods of neuroscience and cognitive development can provide signatures for
tracking the emergence of different systems over development, and testing fundamental questions about developmental continuity.

Recently, some developmentalists have been pessimistic about the contribution of neuroscience to understanding high-level cognition and its development (Keil, 2015). While skepticism may be warranted, I argue that the synergy of these two approaches is far from hopeless. Even focusing on the relatively limited set of findings reported in this thesis, it seems clear that there are the potential advances to be made by combining the perspectives of cognitive development and neuroscience. These approaches do not merely inform where in the brain certain operations occur, or when in development they emerge. Rather, if harnessed effectively, these methods can reveal the representations that underlie emotion perception and inferences, the transformations by which these representations are constructed during online processing, and the mechanisms by which these mechanisms emerge over human development.

4.4 CHALLENGES FOR FUTURE RESEARCH

In this final chapter, I have reviewed three lines of work that shed light on the mechanisms underlying emotion inference, and have sketched ways in which this kind of research could deepen our understanding of emotion perception and cognition. While the synergy of cognitive development and neuroscience holds tremendous promise, I argue that extant approaches are not without limitations. There are a number of challenges that will remain to be addressed in future research.

First, merging the efforts of developmental science and neuroscience will raise methodological challenges. For example, many of the suggestions described above would depend either on improving methods for studying neural activity noninvasively in young infants (e.g. NIRS, infant fMRI). In some cases, the questions may require very precise spatial resolution, meaning we will need to identify appropriate animal models (see Sereno & Tootell, 2005; Tsao, Moeller, & Freiwald, 2008 on challenges with establishing homologies). When appropriate animal models can be identified, it may be particularly fruitful to combine developmental methods such as controlling rearing (e.g. Sugita, 2008) with the tools of neurophysiology (e.g. Freiwald & Tsao, 2010b;
Freiwald et al., 2009). Moreover, to genuinely inform our understanding of concept acquisition and theory-building, neuroimaging studies will need to move beyond cross-sectional approaches that focus on static snapshots at one or two periods of development (Karmiloff-Smith, 2010). Longitudinal studies that measure change in neural activity over time, and relate those changes to sensitive behavioral measures (e.g. Richardson et al., 2013) may be one promising direction. Training studies that actively intervene on children’s experiences or knowledge and examine consequences for neural organization may be another fruitful approach (see Delazer et al., 2003; Mahy, Moses, & Pfeifer, 2014; Srihasam, Mandeville, Morocz, Sullivan, & Livingstone, 2012).

More importantly, there are key conceptual challenges to be tackled in future work. The research described here generally involved identifying “features” that vary across emotion categories and can explain neural representation of different emotions. As mentioned above, although this feature-based approach has been productive, it is unlikely that representations in all domains, particularly domains of high-level cognition such as theory of mind and emotion inference, can be reduced to operations over lists of associated features (Laurence & Margolis, 1999; Murphy & Medin, 1985). Recently, computational cognitive scientists have formalized theory of mind and its development in terms of probabilistic inference over structured, generative knowledge representations (e.g. rational planning models: Baker et al., 2009). Thus, to capture the richly causal and compositional nature of the representations in these kinds of domains, neuroscientists will need to move beyond a flat, feature vector approach, incorporating the abstract, structured knowledge representations that have been fruitful in other areas of cognitive science (Tenenbaum et al., 2011).

What are the series of processing stages, and what do the representations look like at each step? A feed-forward hierarchy surely isn’t right, but how is a theory encoded? Predictive coding is a hot way to encode theories as far as providing a way to think about predictions and explanations of input data but how are abstract principles or assumptions encoded. What’s hard about high-level cognition is not just that its top-down, but that the constraints and predictions are deeply abstract. Thinking about how brains actually encode principles and assumptions of theories, and use them to generate predictions and constrain interpretations is the golden challenge for cognitive neuroscience.
4.5 FINAL CONCLUSIONS

Despite these challenges, the progress already made in the domain of emotion attribution, and in related areas (e.g. face perception, false-belief reasoning), should leave us optimistic about the approach. Neuroscience allows for characterizing different levels of abstraction, including intermediate stages encapsulated from behavior, and can be used in conjunction with behavioral studies to identify features that govern the representations at different processing stages. This understanding in turn, can be used to understand the innate cognitive architecture that supports the emergence of these representations, and to characterize mechanisms of change in these representations over human development.

Of course, the research reported in this thesis provides only a few modest steps in the direction of such an integrated understanding of human emotion perception and inference. Nonetheless, the present findings are among the first to characterize the neural and developmental basis of emotion inference, rather than mere perceptual recognition, and all three chapters suggest lines of future research that could deepen and expand our understanding of the processes by which we attribute emotions to others. Pulling on methodologies that have been fruitful in the study of domains of visual perception (e.g. face processing, object recognition), the research in this thesis shows that these approaches can be successfully applied to higher-level cognitive processes. Importantly, much of higher-level cognition is characterized by inference over rich, generative models of specific domains, rather than bottom-up, feature-based recognition (Friston, 2002; Tenenbaum et al., 2011; Yuille & Kersten, 2006). A fundamental goal for cognitive neuroscience is therefore to understand how the brain develops abstract, structured representations of the world, and recruits these representations to make inferences in the face of limited data (e.g. extracting information about relationships, goals, and emotions from a few seconds of animated shapes moving about a screen). To make progress towards this goal, we must characterize inferential processes that cannot be solved via a feed-forward hierarchy of filters or feature detectors. Thus, the form of inferential emotion attribution investigated in this thesis could serve as a case study for understanding the mechanisms that support high-level conceptual representation and inference more broadly.
### Table 1. Whole brain, Experiment 1: Searchlight results (p<.05, FWE corrected)

<table>
<thead>
<tr>
<th>Stimulus</th>
<th># Voxels</th>
<th>Peak T</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Situations</td>
<td>52</td>
<td>11.80</td>
<td>4</td>
<td>46</td>
<td>38</td>
<td>DMPFC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8.24</td>
<td>6</td>
<td>50</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>9.49</td>
<td>-8</td>
<td>54</td>
<td>26</td>
<td>MMPFC</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>9.21</td>
<td>4</td>
<td>58</td>
<td>14</td>
<td>MMPFC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.02</td>
<td>4</td>
<td>56</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>7.98</td>
<td>16</td>
<td>60</td>
<td>24</td>
<td>MMPFC</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>7.86</td>
<td>0</td>
<td>50</td>
<td>36</td>
<td>MMPFC</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>7.82</td>
<td>0</td>
<td>54</td>
<td>30</td>
<td>MMPFC</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>7.55</td>
<td>-8</td>
<td>54</td>
<td>18</td>
<td>MMPFC</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7.43</td>
<td>8</td>
<td>56</td>
<td>20</td>
<td>MMPFC</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>7.40</td>
<td>-2</td>
<td>54</td>
<td>36</td>
<td>DMPFC</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>7.30</td>
<td>-28</td>
<td>-78</td>
<td>32</td>
<td>L OCC/TEMP</td>
</tr>
<tr>
<td>Faces</td>
<td>8</td>
<td>8.77</td>
<td>-30</td>
<td>-88</td>
<td>-4</td>
<td>L MID OCC GYRUS</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>8.48</td>
<td>38</td>
<td>-92</td>
<td>8</td>
<td>R MID OCC GYRUS</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>8.16</td>
<td>2</td>
<td>52</td>
<td>20</td>
<td>MMPFC</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>7.88</td>
<td>6</td>
<td>52</td>
<td>22</td>
<td>MMPFC</td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>7.60</td>
<td>8</td>
<td>56</td>
<td>20</td>
<td>MMPFC</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>7.52</td>
<td>28</td>
<td>-82</td>
<td>32</td>
<td>R SUP OCC</td>
</tr>
<tr>
<td>Cross-stimulus</td>
<td>42</td>
<td>10.91</td>
<td>-2</td>
<td>50</td>
<td>34</td>
<td>DMPFC</td>
</tr>
<tr>
<td></td>
<td></td>
<td>9.28</td>
<td>0</td>
<td>48</td>
<td>24</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>7.28</td>
<td>8</td>
<td>56</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>8.93</td>
<td>8</td>
<td>56</td>
<td>10</td>
<td>MMPFC</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>7.34</td>
<td>12</td>
<td>66</td>
<td>10</td>
<td>MMPFC</td>
</tr>
</tbody>
</table>
Table S1. Results from whole brain searchlight for 20-way emotion classification (p<.05, FWE corrected, k>25)

<table>
<thead>
<tr>
<th># voxels</th>
<th>Peak T</th>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>Region</th>
</tr>
</thead>
<tbody>
<tr>
<td>1204</td>
<td>11.91</td>
<td>12</td>
<td>60</td>
<td>18</td>
<td>DMPFC/MMPFC</td>
</tr>
<tr>
<td>527</td>
<td>12.27</td>
<td>50</td>
<td>-58</td>
<td>18</td>
<td>RTPJ</td>
</tr>
<tr>
<td>233</td>
<td>12.99</td>
<td>-46</td>
<td>-66</td>
<td>20</td>
<td>LTPJ</td>
</tr>
<tr>
<td>147</td>
<td>11.40</td>
<td>-46</td>
<td>-56</td>
<td>46</td>
<td>LIPL</td>
</tr>
<tr>
<td>107</td>
<td>9.83</td>
<td>54</td>
<td>2</td>
<td>-22</td>
<td>RSTS</td>
</tr>
<tr>
<td>53</td>
<td>10.55</td>
<td>-58</td>
<td>-18</td>
<td>-16</td>
<td>LSTS/LMTG</td>
</tr>
<tr>
<td>41</td>
<td>9.06</td>
<td>0</td>
<td>-34</td>
<td>32</td>
<td>PC</td>
</tr>
<tr>
<td>36</td>
<td>10.51</td>
<td>-40</td>
<td>24</td>
<td>32</td>
<td>LMFG</td>
</tr>
<tr>
<td>36</td>
<td>9.33</td>
<td>-48</td>
<td>36</td>
<td>12</td>
<td>LIFG</td>
</tr>
<tr>
<td>32</td>
<td>9.27</td>
<td>54</td>
<td>-24</td>
<td>-6</td>
<td>RSTS</td>
</tr>
<tr>
<td>32</td>
<td>8.83</td>
<td>36</td>
<td>48</td>
<td>0</td>
<td>RIFG/MFG</td>
</tr>
<tr>
<td>28</td>
<td>10.39</td>
<td>-2</td>
<td>-18</td>
<td>40</td>
<td>PC</td>
</tr>
<tr>
<td>26</td>
<td>9.40</td>
<td>-50</td>
<td>-36</td>
<td>42</td>
<td>LIPL</td>
</tr>
<tr>
<td>26</td>
<td>8.56</td>
<td>16</td>
<td>52</td>
<td>28</td>
<td>DMPFC</td>
</tr>
</tbody>
</table>
Figure S1. Univariate Analyses of ToM ROIs: We localized theory of mind regions in the majority of subjects using the localizer contrast FB>FP (DMPFC: 20 subjects, MMPFC: 21 subjects, VMPFC: 22 subjects, RTPJ: 22 subjects, LTPJ: 22 subjects, PC: 22 subjects, RSTS: 22 subjects). Previous results suggest that these regions are selectively involved in processing the mental states of other people relative to physical or bodily states (Bruneau et al, 2012; 2013). We confirmed the selectivity of localized ROIs by comparing the response to the emotion stimuli and physical pain stimuli. Plots of percent signal change in each of the ToM ROIs show the BOLD response to the 20 emotion stimuli (negative situations in yellow/green, positive situations in blue/pink) relative to the response to non-mental stimuli describing physical pain (dark red). Consistent with prior work, we found robustly higher response to the emotional items relative to physical pain items (average Emotion beta > Pain beta) in all regions: DMPFC: t(19)=6.224, p<.001; MMPFC: t(20)=6.115, p<.001; VMPFC: t(21)=6.065, p<.001; RTPJ: emotion>pain: t(21)=8.085, p<0.001; LTPJ: t(21)=7.620, p<0.001; PC: t(21)=5.182, p<0.001)
Figure S2. A. Confusion matrix from behavioral stimulus categorizations performed by independent subjects on MTurk. B. Confusion matrices from neural classification analysis (linear SVM trained and tested on 20 emotion categories across runs).
Figure S3. The behavioral classification results show that different emotions have distinct profiles across the 38 appraisal dimensions. Plot shows examples of the feature averages (+/- SEM across items) for several different emotions.
Figure S4. Correlation matrix for 38-dimensional feature space. Pairwise correlations between individual appraisal features show high collinearity amongst certain features.
Figure S5. Reduced appraisal space: A. Stimulus-by-appraisal matrix reconstructed using the 10 selected features, capturing 75% of the variance of the original matrix of stimuli-by-38-dimensions. B. Classification of 20 emotions (generalizing across stimulus exemplars) using information from the full 38-dimensional space (dark blue), the space of 10 selected features (light blue), and each of those 10 features individually (green). A model trained to distinguish the 20 emotion labels using only the 10 features could classify the emotions of novel stimuli at 45% accuracy (compared to 57% observed with the full appraisal space), and the emotional discriminations were not captured by any of the individual features in isolation (all accuracies <20%).
Figure S6. Representational Similarity Analysis Results (secondary ROIs): Mean correlation (kendall’s tau) between model RDMs and individual subject neural RDMs (+/- SEM across subjects). Dotted line shows the correlation of a similarity space defined by the raw behavioral confusion matrix. ROIs were all significantly correlated with the space of 38 appraisals: LTPJ: M(SEM)=0.06(0.02), z(21)=2.97 p<0.001, PC: M(SEM)=0.05(0.01), z(21)=3.07 p<0.001, RSTS: M(SEM)=0.06(0.02), z(20)=2.97 p<0.001, VMPFC: M(SEM)=0.03(0.01), z(17)=1.81 p=0.035.
Figure S7. Representational Similarity Analysis Feature Results (secondary ROIs): Mean correlation between neural RDMs (from LTPJ, RSTS, VMPFC, and PC) and the RDM encoding the 38-dimensional space, an RDM encoding the reduced space of 10 features, and separate RDMs encoding each of the features individually (+/- SEM across subjects).
**Figure S8.** Representational Similarity Time-course Analysis: To explore the temporal profile of representation in each region, RSA analyses were conducted separately for overlapping 4s windows with onsets ranging from 0 to 11 TRs post stimulus presentation. This analysis reveals relatively comparable time-courses across ROIs; for example, the regions with the strongest correlations (DMPFC, MMPFC, and RTPJ) all exhibited a peak in similarity in the window 6-7 TRs post stimulus onset.


Deen, B., Kanwisher, N., & Saxe, R. (in review). Parts-based representations of perceived face movements in the superior temporal sulcus.


Hagan, C. C., Woods, W., Johnson, S., Green, G. G. R., & Young, A. W. (2013). Involvement of right STS in audio-visual integration for affective speech demonstrated using MEG. *PloS One, 8*(8), e70648.


http://doi.org/10.1038/nn.3337


http://doi.org/10.1511/2001.4.344


Repacholi, B. M., Meltzoff, A. N., & Olsen, B. (2008). Infants' understanding of the link between visual perception and emotion: 'If she can't see me doing it, she won't get angry.' *Developmental Psychology, 44*(2), 561–574.


