Linking models of Theory of Mind and measures of human brain activity

Sean Dae Houlihan, Joshua B Tenenbaum, and Rebecca Saxe

Department of Brain and Cognitive Science
Center for Brains, Minds and Machines
McGovern Institute for Brain Research
MIT, Cambridge, MA 02139, United States

1 Abstract

Humans employ a richly structured intuitive theory of psychology to reason about others' unobserved mental states, a faculty called 'Theory of Mind'. Advances in behavioral modeling have begun to capture aspects of the flexible and nuanced reasoning people exhibit when inferring the contents of others' minds. In parallel, advances in neuroimaging have begun to illuminate the structure of neural responses evoked when representing others' minds. Bringing these lines of work together will require precise and testable linking hypotheses about how computations over a causal generative model are implemented by the brain, and how these models are acquired during development. We consider how computational modeling and neuroimaging of Theory of Mind can mutually constrain the space of linking hypotheses.

2 Introduction

Human 'Theory of Mind' includes the abilities to recognize, infer, reason about, respond to, predict, cause and avoid causing specific beliefs, desires and emotions in other people. The central questions for cognitive neuroscience about these abilities, are: (i) How do people compute these inferences online? That is, how do people combine current evidence with structured priors in specific situations to explain what others are thinking, predict what they will do next, or choose how to respond? (ii) How do people learn the structured priors? That is, what combination of evidence, experience, and innate biases drive the acquisition of the framework theory of other minds that people bring to specific interactions? And (iii) How are inference in, and development of, Theory of Mind implemented in the human brain? In this chapter, we consider how existing evidence from human neuroimaging experiments helps to constrain answers to these questions.

Understanding the implementation of Theory of Mind in the brain poses some daunting challenges. Mature human Theory of Mind is likely to be at least partially unique to humans. The human behavioral repertoire of flexible cooperation (including pedagogy) and strategic competition imply that humans have a distinctive kind of social intelligence compared to even our closest primate relatives. To the degree that human Theory of Mind is a function, selectively, of human brains, it raises a methodological problem: the methods that we have to study computation in the brain are dramatically more limited for human brains than for other model systems. All existing non-invasive neuroimaging technologies have limited spatial resolution, temporal resolution, and coverage. Nevertheless, we argue that the harder challenge is not methodological but theoretical.

We will consider some recent attempts to link models of Theory of Mind to measurements of human brains, the advances that these attempts support, the limits of those advances, and some of the possible next steps. Most importantly, we need explicit linking hypotheses, computational models of how dynamics of activity in neural populations could implement inferences in (or learning of) logically and causally structured theories. In this chapter, we will mostly just point to the gaps that future linking hypotheses could potentially fill.

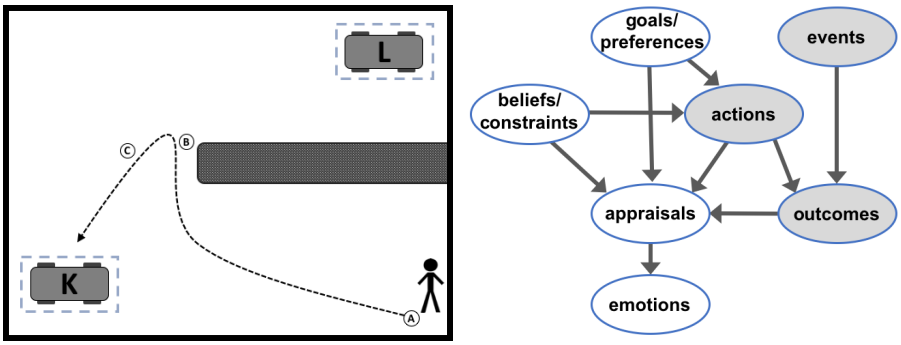
3 How we infer others' mental states

One step in the right direction is to begin with a description of the problem space. What is Theory of Mind, and what is it for? A Theory of Mind is an inferred latent causal structure in another mind. We use Theory of Mind to predict a person's future actions based on our estimates of their unobserved mental states, such as their beliefs (Wimmer and Perner, 1983; Gergely et al., 1995; Jern and Kemp, 2015; Gershman et al., 2016). But Theory of Mind is not only used to anticipate behavior. We explain actions after they occur, changing our understanding of a person's expectations, values, costs, habits and intelligence (Baker et al., 2017; Jara-Ettinger et al., 2016; Jern and Kemp, 2015; Jern et al., 2017; Evans et al., 2016; Gershman et al., 2016; Kryven et al., 2016; Kliemann and Adolphs, 2018). These explanations are themselves value-laden: we use Theory of Mind to make moral judgements of a person's actions and character (Cushman et al., 2013; Kleiman-Weiner et al., 2015). We track a person's knowledge (and its sources) so we know what to learn from her (Gweon and Asaba, 2018; Mills, 2013; Shafto et al., 2012). Our causally structured Theory of Mind shapes how we interpret others' expressions (Anzellotti et al., 2019; Ong et al., 2015; de Melo et al., 2014) and what antecedents evoked them (Wu et al., 2018). We use our intuitive theory of other people's minds to design interventions: to plan how best to teach in order to change others' beliefs (Gweon et al., 2018; Bridgers et al., 2019), or how best to persuade in order to change their desires.

The best known measure of Theory of Mind abilities is the false belief task (Wimmer and Perner, 1983; Schaafsma et al., 2015). In a traditional false belief task, the participant observes a character who forms a belief based on direct perceptual access (e.g. "the ball is in the box"); while the character is no longer present, the reality is altered (e.g. the ball is moved to the basket); and then the observer is asked about the

character's beliefs ("Where does she think the ball is?") or actions ("Where will she first look for the box?"). False belief tasks thus provide a measure of the observer's ability to separately represent where the ball really is, and where the character thinks the ball is.

False belief tasks are useful, but narrow, measures of Theory of Mind; our intuitive causal Theory of Mind supports richer and more generative inferences that include intentions, desires, knowledge, costs, habits, traits and emotions. These inferences are not binary, but continuous and probabilistic, and allow for quantitative variability in performance. Consider, for example, a different task. The participant observes a character (the hungry graduate student Holly) moving around an environment with obstacles (walls) to get reward (lunch) from one of three food trucks (Korean, Lebanese and Mexican). There are two parking spots, so at most two trucks are present on any given day. When Holly leaves her office on this day (point Ⓐ in figure 1a), she can see that the Korean truck is parked in the close southwest space. The Lebanese truck is parked in the far spot in the northeast corner, but she does not know that because the wall is blocking her line of sight. Suppose that she walks past the Korean truck and around to the other side of the wall, where she can now see the Lebanese truck (point Ⓑ). She then turns around and goes back to the Korean truck (point Ⓒ).



(a) (b)

Figure 1: **(a) Holly gets lunch**. From her initial vantage point Ⓐ, Holly can see the Korean food truck (K) in the southwest parking space, but it is not until she reaches point Ⓑ that she can see past the occluding wall to the second parking space. Observers make graded probabilistic attributions of Holly's beliefs, preferences, costs, rewards, prediction errors, counterfactuals, and emotions at every point along her path. **(b) Bayesian Theory of Mind,** depicted here as a directed acyclic graph. Shaded nodes indicate potentially observable variables, open nodes indicate latent variables, and arrows indicate the causal relationship between variables. As this is a model of people's lay theory of other minds, the model's structure, including implied causal relationships, depicts a hypothesis about people's intuitive reasoning, not a scientific hypothesis about the world itself.

To understand Holly's movements, we rely on the central concept of a plan. If her actions are an approximately rational way to achieve her desires given her expectations

and costs, then her actions provide a lot of information about those desires. For example, from the observation that she walked past Korean, saw Lebanese, but selected Korean anyway, observers can infer that Holly prefers Mexican food overall and likes Korean second-best. This is a pretty remarkable inference since observers are systematically inferring Holly's preference for an object that is not present, and that therefore they never observed her choose or even approach. Leveraging this inferred preference, observers can predict Holly's path the next day when the Mexican food truck is parked in the convenient southwestern spot.

In addition to desires, observers can make inferences about Holly's expectations. Because she walked all the way around the building, Holly must have thought it was reasonably likely that the Mexican truck was parked in the northeastern spot. Throughout her path, observers continuously update probabilistic representations of Holly's beliefs and expectations. Inferring Holly's desires and expectations also supports another kind of inference. At the moment she turns the corner and sees the Lebanese truck in the northeastern corner, how does Holly feel? Observers reliably say she feels disappointed: the outcome of her action is going to be less good than she expected (Saxe and Houlihan, 2017).

We can formalize this range of inferences using a probabilistic generative model of Theory of Mind (figure 1b). Observers can estimate Holly's desires, recognize the moment her beliefs change, explain past and predict future actions, and anticipate her emotional reactions. Observers' inferences about Holly are well described by the Bayesian Theory of Mind (BToM) model, which supports inferences about rich latent features by probabilistically inverting a generative model of approximately rational agents perceiving, planning, and acting in a dynamic world (Baker et al., 2017, 2009). To make 'inverse inferences' (inferences of latent mental contents based on the observation causally connect behaviors) of an agent's beliefs and desires by observing its actions, the observer must have priors over the agent's possible beliefs and desires. To understand Holly's search for lunch, we used a flat prior over beliefs (initially agents think all possible world states are equally probable) and a prior about the structure of desires (each agent has a rank ordered preference for the three kinds of food). Starting with these priors, BToM jointly infers an agent's beliefs and desires, conditioned on observing the world state and the agent's actions evolving over time. BToM's inferences match human inferences from these scenarios remarkably well, both quantitatively and qualitatively (Baker et al., 2017). Thus BToM offers a quantitative model of how human observers infer a person's specific beliefs and desires, during the temporal evolution of an event, from observations of the world state and the person's actions.

Since the BToM framework was introduced, a classic line of questions about its interpretation has to do with the nature and origins of the generative model relating beliefs and desires to actions for self versus other. Some theorists who favor a "simulation"-like account of action understanding (e.g., other chapters in this volume) have suggested that BToM provides a computational model of this view, if the generative model of action is taken to be the observer's own action planning mechanism. For independent reasons (Saxe, 2005), we favor a "Theory theory"-like account, where Theory of Mind rests on an intuitive theory or mental model of how agents plan. The generative model is an abstract, compressed representation of the causal structure of minds, likely to be simplified, incomplete, or wrong in various ways, but also applica-

ble in situations that the observer might not themselves have any experience planning in or even be able to plan in. This model could of course be applied to predicting or interpreting one's own actions: people represent their own planning during explicit, conscious intuitive reasoning about one's actions, as in rationalization (Cushman, 2019), and people also have an implicit, unconscious "forward model" of their own planning (McNamee and Wolpert, 2019). An abstract schema of how agents plan in general may contain specific sub-models for one's own planning mechanism as well as the plans of specific well-known individuals. For BToM purposes, probabilities of action sequences in a generative model could be evaluated by various means including but not limited to "simulation-based computations" in the technical engineering sense (e.g., Monte Carlo methods for approximate Bayesian inference). But crucially, none of these possibilities reflect the "simulation" accounts of action understanding that some cognitive theorists have proposed, in that the BToM generative model is not implemented in the observer's own planning mechanisms. Only such an interpretation seems to us consistent with the range of inferences—both successful and unsuccessful—that people can carry out with their intuitive Theory of Mind and that we as scientists can model formally and quantitatively using the BToM framework. Nevertheless, the precise relations between BToM computations applied to one's own versus others' actions and thoughts remains an open question.

A more recent challenge for BToM models is to expand the framework (generally called 'inverse planning' in reference to the inversion of a forward planning model) to more complex and realistic action plans and environments. In the food trucks examples (and related research on lotteries; Ong et al. 2015), a single actor pursues private goals given individual expectations about the physical world. By contrast, Theory of Mind must also apply to understanding actions in pursuit of social goals (including both direct outcomes for others, Kleiman-Weiner et al. 2017a, 2016; Ullman et al. 2009, and the reputation consequences of actions, Kleiman-Weiner et al. 2017b), given expectations that include other people's intentions and actions (Kleiman-Weiner et al., 2016; Jern and Kemp, 2014; Baker et al., 2008; Shum et al., 2019). Incorporating social interactions will also be necessary to capture a wider array of emotion attributions, such as understanding when a character will feel *pride*, *embarrassment* or *envy* (for example Saxe and Houlihan, 2017). Thus, expanding to more naturalistic settings will necessitate learning an appropriate latent space as well as transformations and computations over that space. Behavioral work has pointed to useful primitive functions (e.g. utilities, reward prediction errors, counterfactuals), but the space of possibilities is large. Discovering the representational abstractions made by neural systems involved in Theory of Mind could heavily constrain the hypothesis space and guide complimentary behavioral modeling approaches. One promising approach is probabilistic program induction, where a hierarchical model learns an inductive bias over inverse planning models like BToM (Lake et al., 2015; Ong et al., 2019).

Similarly, a computational model of Theory of Mind should not only match human behavior, but should also suggest hypotheses for neural implementation. We must test how populations of neuronal activity patterns encode the causal structure of another person's inferred expectations, desires and plans. As of now, we still lack any explicit linking hypotheses that could fill this gap. But the results emerging from contemporary neuroimaging experiments suggest we are headed in a useful direction.

4 Neural basis of Theory of Mind inferences

When people are thinking about thinking, a group of brain regions is robustly and reliably recruited (figure 2), including bilateral temporal parietal junction (RTPJ, LTPJ), precuneus (PC), and medial prefrontal cortex (MPFC) (Saxe and Powell 2006; for reviews see Schurz et al. 2014; Saxe and Young 2013; Koster-Hale and Saxe 2013; Spunt et al. 2015). These brain regions, sometimes called the 'Theory of Mind network' show high hemodynamic responses to evocations of characters' mental states, compared to evocations of physical states of the world, in non-linguistic cartoons (Gallagher et al., 2000; Sommer et al., 2007) and movies (Jacoby, Nir et al., 2016), and in stories presented in writing (Fletcher et al., 1995; Saxe and Kanwisher, 2003; Dodell-Feder et al., 2011; Vogeley et al., 2001; Perner et al., 2006; Aichhorn et al., 2009; Spotorno et al., 2012; Mano et al., 2009; Chan and Lavallee, 2015; Feng et al., 2014) or aurally (Bedny et al., 2009; van Ackeren et al., 2012; Hervé et al., 2013), in English (Fletcher et al., 1995; Saxe and Kanwisher, 2003; Dodell-Feder et al., 2011; Bedny et al., 2009), German (Vogeley et al., 2001; Perner et al., 2006; Aichhorn et al., 2009), Dutch (van Ackeren et al., 2012), French (Spotorno et al., 2012; Hervé et al., 2013), Chinese (Feng et al., 2014; Chan and Lavallee, 2015), Japanese (Mano et al., 2009), and American Sign Language (ASL, Richardson et al. 2019). The results from ASL are revealing, because the stimulus (a video of a highly engaging and emotive narrator) is highly social in all conditions; nevertheless activity in this so-called ToM network, in ASL speakers, was high only when the content of the story concerned the mental states of characters. These regions also show much larger responses when thinking about another person's mental states (belief, desires and emotions) than about the internal states of her body (pain, hunger, thirst; Bruneau et al. 2013, 2012; Skerry and Saxe 2015; Saxe and Powell 2006).

Although there is widespread consensus that TPJ, PC and MPFC are all robustly recruited during mental state inference, the question of whether any of these brain regions constitute a domain-specific mechanism for Theory of Mind has remained controversial. There are many subtle shades to this controversy, not all of which will be addressed here. One simple question, however, is whether activity during mental state inference actually reflects a different, domain-general cognitive process, which is just incidentally evoked by tasks requiring Theory of Mind. Many such cognitive processes have been hypothesized (Buckner and Carroll, 2007; Spreng et al., 2009; Lindquist et al., 2012). For example, tasks that require reasoning about others' minds might also typically evoke rich episodic memories of one's own similar experiences. Episodic memories do evoke activity in a group of brain regions with a similar distribution across cortex, resembling the so-called 'default mode network' (DMN; e.g. Yeo et al. 2011; Fox et al. 2005; Raichle et al. 2001). However, activation of episodic memories cannot explain away the activity in Theory of Mind tasks, because upon closer examination, episodic memory and Theory of Mind recruit activity in almost completely non-overlapping (though spatially nearby and interleaved) cortical regions (DiNicola et al., 2019). Standard fMRI methods for data acquisition and analysis blur these neighbouring cortical regions together (Braga et al., 2019; Braga and Buckner, 2017; Wen et al., 2019). By collecting much more data within single participants, and then analyzing individual participants separately to preserve idiosyncratic cortical

anatomy, DiNicola et al. revealed a striking dissociation between 'DMN's: one involved in memory and projection (future oriented thinking), and the other involved in Theory of Mind. Other studies have used similar approaches to differentiate the cortical regions involved in Theory of Mind, from nearby regions involved in detecting unexpected events and shifting attention (Scholz et al., 2009), perceiving facial and vocal expressions of emotion (Deen et al., 2015), and recognizing social interactions (Isik et al., 2017).

Another function that has been proposed for this cortical network, and especially for TPJ, is narrative comprehension. Responses in TPJ are most robust when a character's mental state is described or evoked in the context of a larger, coherent narrative (Lin et al., 2018). When the narrative coherence is broken, for example by scrambling sentences from a story or scenes from a movie, the response in TPJ is dramatically reduced (Lin et al., 2018; Lerner et al., 2011; Hasson et al., 2008). An explicit statement of a character's mental state (e.g."Sarah believes that swimming in the pool is a good way to get cool"), presented in isolation, does not in fact evoke a very strong response in TPJ; narrative context strongly amplifies these regions' response to the same element. An interesting puzzle is therefore how to understand the cognitive and neural dependency between narrative comprehension and Theory of Mind (Jacoby and

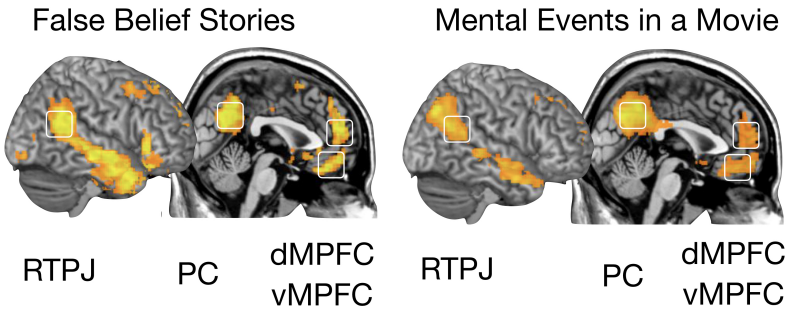


Figure 2: **Thinking about thinking: brain regions commonly recruited in Theory of Mind tasks.** (Left) Average activation in adults reading stories about others' false beliefs (mental state inference), compared to reading stories about false photographs (non-mental conditions that also requires subjects to represent false or outdated content, e.g., an old photograph that no longer accurately depicts the landscape), overlaid on a template brain (Dodell-Feder et al., 2011; Saxe and Kanwisher, 2003). (Right) Average activation in adults watching a Pixar animated short film (*Partly Cloudy*), at the moments of salient mental events (e.g. social rejection/isolation, a baby crying and then becoming happy), compared to salient physical events (slapstick physical harm including the protagonist being poked by porcupine quills or bitten by a baby alligator), overlaid on a template brain (Richardson et al., 2018; Jacoby, Nir et al., 2016). RTPJ: right temporo-parietal junction; PC: precuneus; vMPFC: ventral Medial Prefrontal Cortex; dMPFC: dorsal Medial Prefrontal Cortex. Here we collectively term these cortical regions the Theory of Mind network; they are also known as the Mentalizing network.

Fedorenko, 2018; Schurz et al., 2014; Mar, 2011)). Is there a cortical network for narrative comprehension, that is typically evoked in ToM tasks but might also be evoked when representing any coherent sequences of events or sentences? Or is there a cortical network for Theory of Mind, which is more robustly recruited when mental states are presented in a coherent narrative context? Although these hypotheses have not been definitively tested, evidence favours the latter interpretation. Coherent expository texts with no mental state content evoke minimal responses in Theory of Mind brain regions (Jacoby, Nir et al., 2016; Dodell-Feder et al., 2011); and scrambling these texts has no effect on responses in TPJ (Jacoby and Fedorenko, 2018). Temporally scrambling naturalistic movies (i.e. feature films and TV episodes) does dramatically alter activity in TPJ, but of course these films are designed to evoke rich understanding of characters' minds. Scrambling the order of events plausibly impairs participants' ability to understand and represent the character's more subtle beliefs, desires and emotions.

The Theory of Mind network is thus a set of cortical regions where activity is robustly and selectively evoked by consideration of people's minds. Just finding that a region is selectively active does not address the cognitive or computational questions we posed at the beginning of the chapter. What role do these cortical regions play during online Theory of Mind inferences? One way to investigate is to adapt an approach that has proved highly successful for the ventral visual stream, which is involved in object recognition. A visual image is represented in distinct formats across cortical areas in the ventral visual stream. Low-level stimulus properties like line orientation and shading are linearly decodable from small populations of neurons in early visual areas (e.g., V1) whereas in higher-level regions, the identity of an object becomes linearly decodable and invariant across viewing conditions (DiCarlo et al., 2012; Kamitani and Tong, 2005; Kourtzi and Kanwisher, 2001; Lafer-Sousa and Conway, 2013; Tanaka, 1993). As information propagates through the ventral pathway, the neural response is reformatted to make features that are relevant to object identity more explicit. Discovering which features of a stimulus can be linearly decoded from each population of neurons can reveal the kinds of representations that those populations support.

By analogy to the visual system, we can ask what features of inferred mental states can be linearly decoded from the patterns of activity in cortical regions. Perhaps amazingly, within Theory of Mind brain regions, different spatial patterns of activity are reliably evoked by descriptions of subtly different mental states, so multi-voxel pattern analyses (MVPA) can be used to find meaningful feature dimensions in the patterns of neural responses to others' mental states. For example, as a first proof of principle, we tested whether patterns of activity in RTPJ differentiate representations of an agent knowingly or unknowingly causing harm. How much a person is blamed for a harmful action (e.g. putting poison in a drink, failing to help someone who is hurt, making an insensitive remark) depends substantially on whether the person reasonably believed that her action would (or would not) cause harm. This aspect of moral evaluation depends disproportionately on the function of RTPJ: causally interfering with activity in the RTPJ shifts moral judgments away from reliance on mental states (Young et al., 2010). Spatial patterns of activity in RTPJ (i.e. which subsets of voxels are relatively more, or relatively less active, within this one region) reliably depend on, and therefore can be used to decode, whether a harmful action was taken with full foreknowledge versus in ignorance. Moreover, individual differences in moral judgment were predicted by

individual differences in neural pattern confusability in the RTPJ: people whose RTPJ showed more differentiated patterns of response to intentional vs accidental harms also assigned less blame and greater permissibility to justified accidents (Koster-Hale et al., 2013).

Subsequent research has revealed that the distinction between knowing and unknowing harm is one of many distinctions relevant to Theory of Mind inferences that are decodable from patterns of activity in ToM regions (Koster-Hale et al., 2014, 2013, 2017; Skerry and Saxe, 2015, 2014; Tamir et al., 2016). The clearest distinction between mental states, based on the patterns of activity evoked in ToM regions, is the valence (or goal-congruence) of the state: did the person get (or expect to get) what she wanted? Although valence is an organizing dimension of all Theory of Mind regions, the representation of this dimension appears to depend disproportionately on MPFC function (Amodio and Frith, 2006; Etkin et al., 2011; Hynes et al., 2006; Sebastian et al., 2012; Shamay-Tsoory, 2011; Shamay-Tsoory and Aharon-Peretz, 2007; Shamay-Tsoory et al., 2006; Leopold et al., 2012). The population-level activity in MPFC contains abstract, multimodal information about the valence of another person's experience (Chavez and Heatherton, 2014; Chib et al., 2009; Chikazoe et al., 2014; Kable and Glimcher, 2007; Winecoff et al., 2013). For example, how pleasant the experience is for the protagonist (i.e., the valence of the experience) best explains the pattern of response in MPFC to verbal descriptions of 200 unique emotional events (Skerry and Saxe, 2015). Furthermore, distinct patterns of activity in MPFC are evoked when observing another person (a) make a positive versus negative dynamic facial expression (Harry et al., 2013; Peelen et al., 2010; Said et al., 2010b,a), (b) make a positive versus negative vocal expression (Peelen et al., 2010), (c) succeed versus fail to complete a goal (like throwing a ball into a net) (Skerry and Saxe, 2014), or (d) get included in versus excluded from a social group (Skerry and Saxe, 2014). This diverse range of stimuli evokes a common multivariate representation of valence such that a linear classifier trained to decode valence based on stimuli from one domain (e.g. stereotypical positive and negative facial expressions) was able to decode valence in a different domain (e.g. animations of expressionless shapes succeeding and failing to accomplish goals) (Skerry and Saxe, 2014).

In addition to distinctions relevant to goals, there are also distinctions relevant to plans or beliefs — including distinctions between planned and unplanned states, and between justified and unjustified beliefs: that is, epistemic features. As in the example of Holly above, observers keep sensitive track of others' expectations, including when and how beliefs change through perception and through inference. RTPJ appears to be differentially important for evaluating other people's beliefs and motivations. The features of another's mind that can be decoded from patterns of activity in RTPJ are epistemic: aspects of the inferred process by which she formed her beliefs. These features include properties of her evidence (e.g. whether her source was something she saw or something she heard; Koster-Hale et al. 2014) and properties of the inference process itself (e.g. whether her conclusions were justified by her evidence or not; Koster-Hale et al. 2017). Evidence justification provides a particularly strong test for features of intuitive epistemology because it is abstract (rather than tied to specific sensory features), context specific (what might be good evidence for one conclusion could be poor evidence for another), and directly related to reasoning about the minds

of others (determining whether the agent is a reliable, rational informant; Kovera et al. 1991; Miene et al. 1993; Olson 2003).

As an aside, this distinction between motivation- and valence-biased representations in MPFC, and epistemic representations in RTPJ, may help to resolve a puzzle in the cognitive neuroscience of morality. When a protagonist is described as causing harm knowingly versus unknowingly (e.g. you absolutely knew, versus had no idea about, your cousin's allergy when you served him the peanuts), distinct patterns of activity were observed in RTPJ, and predicted participants' moral judgments of the protagonist (Koster-Hale et al., 2013). By contrast, in a separate experiment, ventral MPFC activity was selective for harmful actions depicted as intentional versus accidental (e.g. deliberately pushing someone versus tripping and falling against them) (Decety et al., 2012). Furthermore, developmental increase in ventral MPFC selectivity for intentional versus accidental harms was associated with developmental reduction in blame for the accidents (Decety et al., 2012). These two sets of results are compatible when viewed in light of the proposed representational architecture for Theory of Mind: RTPJ contains information about what the protagonist knew or should have known, before acting intentionally (i.e. an epistemic feature), whereas the MPFC is sensitive to whether the action was consistent with the protagonist's goals (i.e. a motivational feature).

There are also other distinctions that can be decoded from patterns of activity, for example separating highly social, high arousal states like playfulness, lust, dominance, and embarrassment, from solitary, low-arousal states like exhaustion, laziness, self-pity and relaxation (Tamir et al., 2016). The distinction here may reflect the mental states of others to which we give resource priority—the ones that inspire our urgent attention—because they drive others' actions and demand our own responses. Interestingly, patterns of brain activity in Theory of Mind regions distinguish between justified and unjustified, but not between true and false beliefs (Koster-Hale et al., 2013). These null results are consistent with the argument above: Theory of Mind concerns the process of making rational inferences from perception and knowledge, not whether the beliefs are true or false. Thus, the distinction between true and false beliefs is not given high priority in the neural representations of Theory of Mind. However, null results in MVPA must always be interpreted with caution. Each fMRI voxel potentially contains hundreds of thousands of neurons so many distinct neural populations are intermingled and indistinguishable at this resolution (Freeman et al., 2011; Op de Beeck, 2010).

There are two general lessons of these studies. First, there is remarkable convergence between the cortical locations of peak selective (univariate) responses and peak (multivariate) information, for representations of others' thoughts. The same cortical areas that show the most selective responses to thinking about mental states (i.e. distinguishing mental state information from other conceptual context, between-domains) also contain the most information about mental states (i.e. distinguishing between one type or feature of mental states and another, within-domains). This convergence between evidence of selectivity and evidence of information content strongly suggests that thinking about thought is implemented in domain-specific representational spaces, distinct from other aspects of conceptual and linguistic processing.

Second, and more importantly, pattern analyses have revealed some of the internal structure of mental state representation. These are the observations that should even-

tually allow us to test predictions of alternative computational models of mental state inference. Mental states are not simply represented as different from other kinds of states (of the physical world, of the body), there is also an internal structure of similarity, according to which some inferred mental states elicit more similar patterns of activity in ToM brain regions, and others elicit more distinct patterns. The principal dimensions of this internal structure suggest key divisions of labor within mental state inference.

In summary, fMRI evidence suggests an overall organization of representations of mental states. Others' mental experiences are represented as distinct from their bodily experiences; within concepts of other minds, at least two distinct dimensions are made explicit: one separating positive (goal-congruent) from negative (goal-incongruent) states, and at least one other that may track the source and justification of beliefs.

5 Interpreting computational models in light of neural activity

What do these neuroimaging results reveal about the computations underlying Theory of Mind inferences? One proposal, Tamir and Thornton (2018), is that the similarity structure of brain responses directly reveals the substrate of inferences about minds. Using principal components analysis, they find three main organizing dimensions of activity while participants consider the meaning of 60 different terms for states of mind, ranging from 'anticipation' and 'awe' to 'drunkenness' and 'disarray' (examples given in table 1). Tamir and Thornton (2018) argue that representing other minds in this very low dimensional space explains how people are able to make a key type of inference: prediction. Human observers predict that other people's states of mind are more likely to transition between states that are nearby in this abstract 3D space. For example, we expect that a friend now feeling 'anxious' will be more likely to feel 'sluggish' than 'energetic' later (Thornton and Tamir, 2017). Thus, the predicted dynamics of other minds could be captured by trajectories in a low-dimensional neural representation of types of mental states. This idea is exciting because it is a rare attempt to capture the range and richness of mental state inferences, and because of the explicit linking hypothesis between a neural population code and a cognitive inference mechanism.

We suggest an alternative: that the dynamics of mental states must be understood in terms of causally and logically structured relations between mental contents, not simply transition probabilities. Mental state attributions are not likely to be well-described as simply a list of features; rather, they require representations with internal structure (Baker et al., 2017, 2009; Davidson, 1963), understood in terms of their computational role within a coherent explanatory theory (cf Theory theory; Carey 2009; Gopnik and Wellman 1994). Any representational similarity analysis operationalizes these representations as a "bag of features", more similar 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). This approach contrasts with traditional 'mental states', which are composed of an attitude (or evaluative perception) towards a proposition (or content). We cannot ask how a person's belief will influence her next

Tamir et al.	Skerry and Saxe
Planning: "carrying emergency cash" "executing a science experiment" "looking at the weekend's weather" "researching an item to purchase it" **Belief:** "listening to a religious service" "confident about an attitude" "reading the Bible" "wearing a lucky charm" **Opinion:** "thinking California is the best state" "personal belief" "finding brunettes more attractive" "recommending a type of music" **Thought:** "putting ideas together" "remembering to bring an umbrella" "deciding what to do today" "forming an opinion" **Anticipation:** "on the line to ride a rollercoaster" "waiting for a band to go onstage" **Lust:** "feeling horny" "preferring physical to emotional" **Drunkenness**: "drinking alone" "spending time with an alcoholic"	**Mental:** "Lucy and her teammates trained hard in preparation for the upcoming soccer playoffs. Their coach told them they had a chance of winning the championship. On the first day of the playoffs, a few fluke plays put Lucy's team down 2 to 0. They lost the game, knocking them out of the playoffs in the first round." "Jordan swore to her roommates that she would keep her new diet. Later, she was in the kitchen getting a glass of water, and took a bite of a cake she had bought for their dinner party the following evening. Jordan's roommates arrived home to find that she had eaten half the cake and broken her diet." "Jake always avoided the doctor's office because he really disliked needles. One summer, Jake was traveling to Kenya for a project, and was told he needed a series of tests and vaccinations before he could go. He reluctantly called the travel clinic and scheduled an appointment for the following week." **Physical:** "Roger was walking to school when he heard a friend behind him call his name. Roger turned to respond, but just then tripped and stumbled over some wood on the ground. Roger fell forward and impaled his hand on a rusty nail in the wood."

Table 1: **Example Stimuli:** To capture mental state inferences, Tamir et al. (2016) presented a pair of scenarios and asked participants which would better evoke the associated mental state in another person. For instance, participants indicated whether the mental state "thought" was better evoked by "forming an opinion" or "deciding what to do today". Both scenarios are intended to evoke the associated mental state so there is not a 'correct' answer. Skerry and Saxe (2015) showed participants narratives that prompt mentalistic inferences about plans, beliefs, expectations, desires, reactions and emotions, and narratives that prompt inferences of bodily sensations (these Physical stories do not evoke activity in Theory of Mind regions). Using similar analysis techniques, Tamir et al. concluded that a low-dimensional representational space (4 dimensions) could sufficiently capture behavioral judgements and neural activity during ToM, while Skerry and Saxe concluded that mental state representations are much higher dimensional (> 10 dimensions). One possibility is that Skerry and Saxe's inclusion of richer context, and more specific content, evokes more differentiated cognitive and neural representations of mental states.

action without knowing: her belief about what? Even a simple propositional attitude (e.g. "The father fears his son will fall out of the tree") is composed of an agent (the father), an attitude (fears) and a propositional content (child falling out of tree), and is causally connected to many other specific mental states (e.g. perceptual evidence of wobbly branches, desires to intervene, conflicting desires to promote independence, and so on). The current vector space models do not encode logical or causal structure (context), and lack compositionally (content). The difference between feeling 'playful' versus 'serious' might be measurable as the distance between two vectors along one continuous dimension, but the difference between 'wanting the ball' versus 'wanting to go to the ball', or 'wanting to play' and 'wanting to go to the play', are different in kind. Different formal structures will likely be required (Skerry and Saxe, 2015; Baker et al., 2017). Relatedly, inferences about beliefs necessarily depend on a rich body of world knowledge (e.g. about trees, and about children), so neural populations specific to Theory of Mind must interface with general-purpose semantic systems. A list of features made explicit by each neural population is not enough to test alternative theories of inference in Theory of Mind.

Consistent with this theoretical perspective, there are already empirical hints that representations in the Theory of Mind network are not low-dimensional. We found that patterns of response in the ToM network, including RTPJ and MPFC, can be used to classify verbal narratives (examples given in table 1) into twenty distinct emotion labels (e.g. *furious*, *jealous*, *grateful*, *proud*; Skerry and Saxe 2015). The features that explained significant variance in the neural response are natural components of planning and belief updating, and not all easily captured by the three-dimensional solution: for example, whether the event would be repeated in the future, affected the protagonist's life in the long run, and/or was caused by the protagonist or by other people. We found that a minimum of 10 feature dimensions were required to explain the reliable variance in that dataset, and that is still likely to be a substantial underestimate. Just within the representation of 'rationality' or the reasons for others' beliefs, we have already discovered more than one dimension. In RTPJ, within a single task and set of stimuli, patterns of activity in RTPJ can be used to decode whether the person's beliefs were formed based on sufficient or insufficient evidence, and whether they were based on visual or auditory evidence—the patterns of activity that distinguished beliefs based on modality versus justification were orthogonal (Koster-Hale et al., 2017). Furthermore, evoking rich and specific mental states requires relatively long and complex stimuli. For example,

> *Ginny's classmate wants to borrow a bike to go mountain biking. Ginny's sister left her bike in the garage when she went off to college. The bike had been in and out of the shop for brake trouble. Ginny believes the bike is fully functional now,* since the last time she talked to her sister, the brakes were working fine. *Ginny lends her classmate the bike, which turns out not to be fully fixed. Her classmate crashes into a tree due to the defective brakes and loses her two front teeth.*

implies a justified belief and induces a distinct pattern of activity in RTPJ from the pattern induced by replacing the emphasized text with "though the last time she talked to her sister, the brakes were still giving her trouble". By classifying average neural responses to a whole sentence, presented in the context of a longer narrative, we combined many cognitive processes. As a result, classification results must be

interpreted as a lower bound on the information available in each region (Kriegeskorte and Kievit, 2013).

In sum, we propose that neural populations within the Theory of Mind network support inference by implementing something like the BToM computations: building and operating over a probabilistic causal model of others' motives, expectations and plans. This proposal remains mostly a promissory note. It is missing specific linking hypotheses for how stimuli (i.e. observed events, verbal narratives) are transformed into neural representations, and how priors are represented and combined with representations of the input (which requires a theory of how neurons encode prior knowledge). To make progress in this research program, it will be necessary to construct at least one, but ideally competing, models of how Theory of Mind inference could work in principle, along with more robust linking hypotheses concerning the neural implementation, and the resulting features that might be detectable at the resolution of fMRI. For many reasons, this may fail. But given the current trajectory of progress, it seems worth a shot.

6 Neural basis of Theory of Mind development

A fundamental component of any hypothesis about Theory of Mind inference must be a representation of structured prior knowledge. Holly's movements around her campus can only reveal her preferences and beliefs in virtue of prior knowledge about human planning—that people typically have a rank-ordered preference for foods, that longer paths are more costly, that beliefs can be updated via direct visual access, and so on. How are these priors acquired, and implemented neurally? Using what we know about the mature ToM network, we can operationalize one part of this question by asking how children come to have cortical regions, in RTPJ, MPFC and elsewhere, that are selectively recruited by reasoning about other minds. Is the dramatic and stereotyped development of Theory of Mind abilities during early childhood associated with functional changes in these regions? Are the functions of these brain regions learned? Are they constrained by biological predispositions, and if so, how?

Classic theoretical debates about social cognitive development have considered two opposing possibilities, arguing that ToM is either instantiated in a distinct domain-specific biological mechanism or is constructed through conversational interactions and social relationships (Carlson and Moses, 2001; Scholl and Leslie, 2001; Hughes and Devine, 2015). By contrast, we suggest that ToM is both; Theory of Mind is acquired by a domain-specific biological mechanism, whose mature function and selectivity is constructed in part through linguistically-mediated transmission of culturally-specific concepts.

As described in the previous section, adults have a highly reliable set of cortical regions that are recruited selectively when reasoning about other minds. Activity in these regions is high when thinking about their thoughts or emotions, but not when considering other features of the same characters, including their physical actions and bodily sensations. We argued earlier that these regions constitute a domain-specific biological mechanism, with a selective function in Theory of Mind. The functions of these regions are distinct from other aspects of social cognition very early in development.

In three year old children, before they can pass false belief tasks, the ToM regions are functionally correlated with each other and respond to evocations of characters' mental states (Richardson et al., 2018). Activity in RTPJ peaks when characters have a false belief, even in preverbal infants (Hyde et al., 2018). Thus, in some sense ToM regions are predisposed to some function related to Theory of Mind, from very early in development. These early origins are not incompatible with environmental influence. On the contrary, we hypothesize that the specific representations and computations of these regions are shaped during development through conversational interactions and social relationships.

Activity in the RTPJ is particularly selective for thinking about others' thoughts in adults (Saxe and Powell, 2006; Mitchell et al., 2005; Saxe and Kanwisher, 2003; Jacoby, Nir et al., 2016; Dodell-Feder et al., 2011; Bruneau et al., 2012; Spunt et al., 2015; Lombardo et al., 2010). Similar to the development of cortical regions specialized for other functions, the development of increased selectivity in the RTPJ occurs by the suppression of responses to non-preferred stimuli. For example, selectivity of the fusiform face area (FFA) develops through the suppression of responses to (non-preferred) non-face objects; this suppression is correlated with performance on face recognition tasks (Cantlon et al., 2010; Golarai et al., 2007; Gomez et al., 2017). Selectivity of the visual word form area (VWFA) develops through the suppression of responses to (non-preferred) faces (Cantlon et al., 2010), and this suppression predicts literacy and reading expertise (Dehaene et al., 2010; Dehaene-Lambertz et al., 2018). Similarly, selectivity of the RTPJ develops through suppression of responses to other (non-mentalistic) social information (Gweon et al., 2012; Saxe et al., 2009), and correlates with performance on ToM tasks (Gweon et al., 2012). For example, in adults, verbal descriptions of a person's physical appearance, place of origin, or social relationships elicit little activity in RTPJ, compared to descriptions of a person's beliefs, desires and emotions (Saxe and Powell, 2006; Mitchell et al., 2005; Gweon et al., 2012). In young children, all of these different kinds of social cues evoke high responses in RTPJ, compared to non-social controls (e.g. descriptions of the physical environment) (Gweon et al., 2012; Saxe et al., 2009).

In the case of FFA and VWFA, extensive domain-relevant experience precedes the emergence of a selective cortical region. What drives the developmental acquisition of RTPJ selectivity and what role does environmental experience play? A particularly important source of input that children use to build a Theory of Mind is linguistically rich conversational experience. In conversation, adults use words and sentences to describe their mental states and experiences (Harris, 2002, 1992). Even utterances that do not contain mental state verbs (e.g. "Where is my hat?") frequently provide evidence about another person's beliefs and desires, which then help to interpret behavior (Siegal and Peterson, 1994; Peterson and Siegal, 2016). However, utterances that do include mental state verbs may be a particularly rich source of information: children learn to differentiate mental state concepts (e.g. believe vs. know, want vs. hope, peek vs stare) from the way adults use these mental state verbs in conversational context (Gleitman, 1990). Indeed, just the existence of these distinct words may be an important source of evidence to children, concerning the structure and kinds of mental state concepts used in their culture.

The clearest evidence that linguistic experience affects ToM development comes

from studies of children who are d/Deaf and not exposed natively to a sign language. Many deaf or hard of hearing children are at risk of not learning any language in early childhood because they have limited auditory access to spoken language, and their families do not know sign language at the time of birth (Mitchell and Karchmer, 2004). Deaf children with delayed exposure to sign language show corresponding delays in ToM relative to typically hearing children and deaf children exposed to sign language from infancy (Peterson and Siegal, 2016; Woolfe et al., 2002; Schick et al., 2007; Gale et al., 1996; Schick and Hoffmeister, 2001; Figueras-Costa and Harris, 2001; Peterson et al., 2005, 2012; Peterson and Wellman, 2018). Hearing parents who learn sign language as a second language exhibit large variability in their use of mental state language, which in turn predicts their deaf children's performance on ToM tasks (Moeller and Schick, 2006).

We therefore tested the effect of delayed access to language on the development of selectivity in RTPJ (Richardson et al., 2019). In native signing children, the RTPJ showed selective responses to stories about mental states in the linguistic ToM task. Like native signers, delayed signing children showed high responses to Mental stories ("Jimmy soon realized the pirate didn't know where the treasure was"), but the response in their RTPJ was also high for non-mentalistic social information – like physical appearances and enduring relationships (Social stories: "Old Mr. McFeegle is a gray wrinkled old farmer"; "Sarah and Lori play together on the soccer team"). The reduced selectivity in RTPJ was similar to the response profile previously observed in young children (Gweon et al., 2012; Saxe et al., 2009). Delayed access to ASL correlated with delayed selectivity of RTPJ for mental state information, despite relatively short delays prior to language exposure, and despite being highly proficient in ASL comprehension (matched to native signers) at the time of testing (Richardson et al., 2019).

Conversational experience is not only necessary for acquisition of mental state concepts, it can also be sufficient. The clearest evidence for the sufficiency of conversational exposure comes from the incredible richness of congenitally blind people's knowledge about sight. If first-person experience is necessary to understand others' experiences, blind people should have only a fragmentary, limited, or metaphorical understanding of seeing. But they don't. On the contrary, through conversation and social interaction with sighted people, blind people acquire a rich intuitive theory of sight. Even young blind children know that other people can see with their eyes, and understand for example that objects can be seen from a distance and are invisible in the dark (Bigelow, 1992; Landau and Gleitman, 1985; Peterson et al., 2000). By adulthood, congenitally blind people know the meanings of verbs of sight, including fine-grained distinctions between concepts like peer, gaze, and gawk (Bedny et al., 2019; Landau and Gleitman, 1985; Lenci et al., 2013). Finally, the similarity between blind and sighted people's reasoning about sight is evident not just in behavior but also in neural implementation. Like sighted people, blind people recruit RTPJ selectively when thinking about other people's experiences of seeing, but not their experience of bodily states like hunger or nausea (Bedny et al., 2009), and the pattern of neural activity in the RTPJ of both blind and sighted people can decode the source of the character's belief from auditory vs visual evidence (Koster-Hale et al., 2014).

In summary, we propose that during development, children learn a model of the

latent causal structure of other minds. This learning occurs through conversational interactions and social relationships, and thus is attuned to the distinctions and structures of other minds that are relevant in the child's cultural context. On the other hand, learning some kind of model of other minds is in a sense biologically prepared by, and preferentially attached to, a reliable cortical mechanism and thus appears in the same highly selective regions across individuals, languages and groups. What is learned by these cortical regions must be not only a division of the domain of minds from other aspects of social life, but also the structured priors (i.e. the framework theory) about how minds work in general that supports specific inferences about one person's beliefs or desires in one particular context. As above, future work is required to define testable linking hypothesis for how development of domain-specific brain regions constitutes the construction of structured priors for inferences.

7 Future directions: linking neural measures to computational models

For the next step in a deeper understanding of both inference and development of Theory of Mind, we need well-specified hypotheses for how neural dynamics could implement computations over a mental model of latent causal structure. This is a lofty goal, and not unique to Theory of Mind. Other domains of cognitive neuroscience, including the neural basis of language and of intuitive physics, face a similar challenge. The solution to this challenge is unknown, so here we point in some promising future directions.

The first step is to define a range of Theory of Mind inferences that (i) covers the rich and elaborated structure of the intuitive Theory of Mind, and (ii) can be well captured by computational models of inferences. We propose that a good starting point is inferences about others' reactions to unfolding events (Saxe and Houlihan, 2017; Ong et al., 2015). Predicting another person's reactions requires a causal model of their mind, because reactions happen when people's expectations, desires, plans and habits meet a dynamic world. For example, when Holly the graduate student sets out looking for lunch, her plans reveal her expectations (where the food trucks will be) and preferences (which cuisines she prefers). At the moment that she turns the corner and sees her least favourite truck parked in the northeast spot, observers infer that Holly can update her expectations based on her perception (an epistemic change). Because changing her expectations about the trucks changes her expected reward in the situation, observers also recognize that Holly is experiencing negative reward prediction error—that is, disappointment (Ong et al., 2015; Wu et al., 2018).

We propose that BToM can be expanded to match human observers' inference about others' emotions (Saxe and Houlihan, 2017). BToM probabilistic generative models are designed to update posterior estimates of a person's preferences and expectations based on her actions, and then compute the consequences of events in terms of the person's achieved utilities (did she get what she wanted), prediction errors (did she get what she expected), counterfactual utilities (what would she have gotten if she chose a different action), and so on. If, as we suggest, these features are core compo-

nents of Theory of Mind inferences, then they should also provide a good fit to neural activity during those inferences (Skerry and Saxe, 2015). That is, the features computed by BToM could be used as an encoding model (Mitchell et al., 2008; Naselaris et al., 2011) for fMRI responses: an explicit hypothesis about the features represented explicitly in the Theory of Mind brain regions.

8 Conclusions

How are inference in, and development of, Theory of Mind, implemented in the human brain? Here we argue that Theory of Mind inferences are implemented, at least partially, in distinct and selective cortical regions. Within these regions, neural activity is generally high and sustained, while people think about thoughts, and distinct patterns of population activity contain information about abstract dimensions or features of the inferred mental states, including valence and rationality. The strong selectivity, and presumably the distinct spatial patterns, in these cortical regions emerge reliably during development. However, adult cortical divisions of labour are not fully innately prespecified, but rather emerge in social and cultural context. As yet, there are no testable (let alone competing) models linking the activity in these cortical regions to adequate inferential processes over causal models, that can capture the sensitivity of human Theory of Mind. Development of such models is a critical direction for future research.

8.1 Acknowledgments

We would like to sincerely thank Hilary Richardson for providing figure 2. We are grateful for the immensely useful feedback we received from Daniel Nettle, Stefano Anzellotti and Michael Gilead on an earlier draft of this chapter. This work was supported by the McGovern Institute and the Center for Brains, Minds and Machines (CBMM; funded by NSF STC award CCF-1231216).

References

Aichhorn, M., Perner, J., Weiss, B., Kronbichler, M., Staffen, W., and Ladurner, G. (2009). Temporo-parietal Junction Activity in Theory-of-Mind Tasks: Falseness, Beliefs, or Attention. *Journal of cognitive neuroscience*, 21(6):1179–1192.

Amodio, D. M. and Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature reviews Neuroscience*, 7(4):268–277.

Anzellotti, S., Houlihan, S. D., Liburd, S., and Saxe, R. (2019). Leveraging facial expressions and contextual information to investigate opaque representations of emotions. *Emotion (Washington, D.C.)*.

Baker, C. L., Goodman, N. D., and Tenenbaum, J. B. (2008). Theory-based Social Goal Inference. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*.

Baker, C. L., Jara-Ettinger, J., Saxe, R., and Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, 1(4):598.

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

Bedny, M., Koster-Hale, J., Elli, G., Yazzolino, L., and Saxe, R. (2019). There's more to "sparkle" than meets the eye: Knowledge of vision and light verbs among congenitally blind and sighted individuals. *Cognition*, 189:105–115.

Bedny, M., Pascual-Leone, A., and Saxe, R. R. (2009). Growing up blind does not change the neural bases of Theory of Mind. *Proceedings of the National Academy of Sciences of the United States of America*, 106(27):11312–11317.

Bigelow, A. E. (1992). Blind children's ability to predict what another sees. *Journal of Visual Impairment & Blindness*, 86(4):181–184.

Braga, R. M. and Buckner, R. L. (2017). Parallel Interdigitated Distributed Networks within the Individual Estimated by Intrinsic Functional Connectivity. *Neuron*, 95(2):457–471.e5.

Braga, R. M., Van Dijk, K. R. A., Polimeni, J. R., Eldaief, M. C., and Buckner, R. L. (2019). Parallel distributed networks resolved at high resolution reveal close juxtaposition of distinct regions. *Journal of Neurophysiology*, 121(4):1513–1534.

Bridgers, S., Jara-Ettinger, J., and Gweon, H. (2019). Young children consider the expected utility of others' learning to decide what to teach. *Nature Human Behaviour*, 16:382.

Bruneau, E., Dufour, N., and Saxe, R. (2013). How We Know It Hurts: Item Analysis of Written Narratives Reveals Distinct Neural Responses to Others' Physical Pain and Emotional Suffering. *PLoS ONE*, 8(4):1–9.

Bruneau, E. G., Pluta, A., and Saxe, R. (2012). Distinct roles of the 'Shared Pain' and 'Theory of Mind' networks in processing others' emotional suffering. *Neuropsychologia*, 50(2):219–231.

Buckner, R. L. and Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, 11(2):49–57.

Cantlon, J. F., Pinel, P., Dehaene, S., and Pelphrey, K. A. (2010). Cortical representations of symbols, objects, and faces are pruned back during early childhood. *Cerebral cortex (New York, N.Y. : 1991)*, 21(1):191–199.

Carey, S. (2009). *The Origin of Concepts*. Oxford University Press.

Carlson, S. M. and Moses, L. J. (2001). Individual Differences in Inhibitory Control and Children's Theory of Mind. *Child development*, 72(4):1032–1053.

Chan, Y.-C. and Lavallee, J. P. (2015). Temporo-parietal and fronto-parietal lobe contributions to theory of mind and executive control: an fMRI study of verbal jokes. *Frontiers in psychology*, 6:1405.

Chavez, R. S. and Heatherton, T. F. (2014). Multimodal frontostriatal connectivity underlies individual differences in self-esteem. *Social Cognitive and Affective Neuroscience*, 10(3):364–370.

Chib, V. S., Rangel, A., Shimojo, S., and O'Doherty, J. P. (2009). Evidence for a Common Representation of Decision Values for Dissimilar Goods in Human Ventromedial Prefrontal Cortex. *Journal of Neuroscience*, 29(39):12315–12320.

Chikazoe, J., Lee, D. H., Kriegeskorte, N., and Anderson, A. K. (2014). Population coding of affect across stimuli, modalities and individuals. *Nature Neuroscience*, 17(8):1114–1122.

Cushman, F. (2019). Rationalization is rational. *The Behavioral and brain sciences*, pages 1–69.

Cushman, F., Sheketoff, R., Wharton, S., and Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, 127(1):6–21.

Davidson, D. (1963). Actions, reasons, and causes. *The journal of philosophy*, 60(23):685–700.

de Melo, C. M., Carnevale, P. J., Read, S. J., and Gratch, J. (2014). Reading people's minds from emotion expressions in interdependent decision making. *Journal of personality and social psychology*, 106(1):73–88.

Decety, J., Michalska, K. J., and Kinzler, K. D. (2012). The contribution of emotion and cognition to moral sensitivity: A neurodevelopmental study. *Cerebral cortex (New York, N.Y. : 1991)*, 22(1):209–220.

Deen, B., Koldewyn, K., Kanwisher, N., and Saxe, R. (2015). Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. *Cerebral cortex (New York, N.Y. : 1991)*, 25(11):4596–4609.

Dehaene, S., Pegado, F., Braga, L. W., Ventura, P., Nunes Filho, G., Jobert, A., Dehaene-Lambertz, G., Kolinsky, R., Morais, J., and Cohen, L. (2010). How learning to read changes the cortical networks for vision and language. *Science (New York, NY)*, 330(6009):1359–1364.

Dehaene-Lambertz, G., Monzalvo, K., and Dehaene, S. (2018). The emergence of the visual word form: Longitudinal evolution of category-specific ventral visual areas during reading acquisition. *PLoS Biology*, 16(3):e2004103.

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3):415–434.

DiNicola, L. M., Braga, R. M., and Buckner, R. L. (2019). Parallel Distributed Networks Dissociate Episodic and Social Functions Within the Individual. *bioRxiv*, page 733048.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., and Saxe, R. (2011). fMRI item analysis in a theory of mind task. *NeuroImage*, 55(2):705–712.

Etkin, A., Egner, T., and Kalisch, R. (2011). Emotional processing in anterior cingulate and medial prefrontal cortex. *Trends in Cognitive Sciences*, 15(2):85–93.

Evans, O., Stuhlmüller, A., and Goodman, N. D. (2016). Learning the preferences of ignorant, inconsistent agents. In *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, pages 323–329. University of Oxford, Oxford, United Kingdom.

Feng, S., Ye, X., Mao, L., and Yue, X. (2014). The activation of theory of mind network differentiates between point-to-self and point-to-other verbal jokes: an fMRI study. *Neuroscience letters*, 564:32–36.

Figueras-Costa, B. and Harris, P. (2001). Theory of Mind Development in Deaf Children: A Nonverbal Test of False-Belief Understanding. *The Journal of Deaf Studies and Deaf Education*, 6(2):92–102.

Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S. J., and Frith, C. D. (1995). Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition*, 57(2):109–128.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., and Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9673–9678.

Freeman, J., Brouwer, G. J., Heeger, D. J., and Merriam, E. P. (2011). Orientation Decoding Depends on Maps, Not Columns. *Journal of Neuroscience*, 31(13):4792–4804.

Gale, E., De Villiers, P., De Villiers, J., and Pyers, J. (1996). Language and theory of mind in oral deaf children. In *Proceedings of the 20th annual Boston University conference on language development*, pages 213–224. Cascadilla Press.

Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., and Frith, C. D. (2000). Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia*, 38(1):11–21.

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

Gershman, S. J., Gerstenberg, T., Baker, C. L., and Cushman, F. A. (2016). Plans, Habits, and Theory of Mind. *PLoS ONE*, 11(9):e0162246–24.

Gleitman, L. (1990). The Structural Sources of Verb Meanings. *Language Acquisition*, 1(1):3–55.

Golarai, G., Ghahremani, D. G., Whitfield-Gabrieli, S., Reiss, A., Eberhardt, J. L., Gabrieli, J. D., and Grill-Spector, K. (2007). Differential development of high-level visual cortex correlates with category-specific recognition memory. *Nature Neuroscience*, 10(4):512.

Gomez, J., Barnett, M. A., Natu, V., Mezer, A., Palomero-Gallagher, N., Weiner, K. S., Amunts, K., Zilles, K., and Grill-Spector, K. (2017). Microstructural proliferation in human cortex is coupled with the development of face processing. *Science (New York, NY)*, 355(6320):68–71.

Gopnik, A. and Wellman, H. M. (1994). The theory theory. *Mapping the mind: Domain specificity in cognition and culture*, page 257.

Gweon, H. and Asaba, M. (2018). Order Matters: Children's Evaluation of Underinformative Teachers Depends on Context. *Child development*, 89(3):e278–e292.

Gweon, H., Dodell-Feder, D., Bedny, M., and Saxe, R. (2012). Theory of mind performance in children correlates with functional specialization of a brain region for thinking about thoughts. *Child development*, 83(6):1853–1868.

Gweon, H., Shafto, P., and Schulz, L. (2018). Development of children's sensitivity to overinformativeness in learning and teaching. *Developmental Psychology*, 54(11):2113–2125.

Harris, P. L. (1992). From simulation to folk psychology: The case for development. *Mind & Language*, 7(1-2):120–144.

Harris, P. L. (2002). What do children learn from testimony? In Carruthers, P., Stich, S., and Siegal, M., editors, *The cognitive basis of science*, pages 316–334. Cambridge Univ Pr, Cambridge.

Harry, B., Williams, M. A., Davis, C., and Kim, J. (2013). Emotional expressions evoke a differential response in the fusiform face area. *Frontiers in human neuroscience*, 7:692.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., and Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *The Journal of neuroscience*, 28(10):2539–2550.

Hervé, P.-Y., Razafimandimby, A., Jobard, G., and Tzourio-Mazoyer, N. (2013). A Shared Neural Substrate for Mentalizing and the Affective Component of Sentence Comprehension. *PLoS ONE*, 8(1):e54400.

Hughes, C. and Devine, R. T. (2015). Individual Differences in Theory of Mind From Preschool to Adolescence: Achievements and Directions. *Child Development Perspectives*, 9(3):149–153.

Hyde, D. C., Simon, C. E., Ting, F., and Nikolaeva, J. I. (2018). Functional Organization of the Temporal-Parietal Junction for Theory of Mind in Preverbal Infants: A Near-Infrared Spectroscopy Study. *The Journal of neuroscience*, 38(18):4264–4274.

Hynes, C. A., Baird, A. A., and Grafton, S. T. (2006). Differential role of the orbital frontal lobe in emotional versus cognitive perspective-taking. *Neuropsychologia*, 44(3):374–383.

Isik, L., Koldewyn, K., Beeler, D., and Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences of the United States of America*, 114(43):E9145–E9152.

Jacoby, N. and Fedorenko, E. (2018). Discourse-level comprehension engages medial frontal Theory of Mind brain regions even for expository texts. *Language, Cognition and Neuroscience*, 0(0):1–17.

Jacoby, Nir, Bruneau, Emile, Koster-Hale, Jorie, and Saxe, Rebecca (2016). Localizing Pain Matrix and Theory of Mind networks with both verbal and non-verbal stimuli. *NeuroImage*, 126:39–48.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., and Tenenbaum, J. B. (2016). The Naïve Utility Calculus: Computational Principles Underlying Commonsense Psychology. *Trends in Cognitive Sciences*, 20(8):589–604.

Jern, A. and Kemp, C. (2014). Reasoning about social choices and social relationships. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*.

Jern, A. and Kemp, C. (2015). A decision network account of reasoning about other people's choices. *Cognition*, 142:12–38.

Jern, A., Lucas, C. G., and Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, 168:46–64.

Kable, J. W. and Glimcher, P. W. (2007). The neural correlates of subjective value during intertemporal choice. *Nature Neuroscience*, 10(12):1625–1633.

Kamitani, Y. and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, 8(5):679–685.

Kleiman-Weiner, M., Gerstenberg, T., Levine, S., and Tenenbaum, J. B. (2015). Inference of Intention and Permissibility in Moral Decision Making. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.

Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., and Tenenbaum, J. B. (2016). Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Kleiman-Weiner, M., Saxe, R., and Tenenbaum, J. B. (2017a). Learning a commonsense moral theory. *Cognition*, 167:107–123.

Kleiman-Weiner, M., Shaw, A., and Tenenbaum, J. B. (2017b). Constructing Social Preferences From Anticipated Judgments: When Impartial Inequity is Fair and Why? In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*.

Kliemann, D. and Adolphs, R. (2018). The social neuroscience of mentalizing: challenges and recommendations. *Current Opinion in Psychology*, 24:1–6.

Koster-Hale, J., Bedny, M., and Saxe, R. (2014). Thinking about seeing: Perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults. *Cognition*, 133(1):65–78.

Koster-Hale, J., Richardson, H., Velez, N., Asaba, M., Young, L., and Saxe, R. (2017). Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs. *NeuroImage*, 161:9–18.

Koster-Hale, J. and Saxe, R. (2013). Functional neuroimaging of theory of mind. In Baron-Cohen, S., Tager-Flusberg, H., and Lombardo, M. V., editors, *Understanding other minds*, pages 132–163. Oxford university press.

Koster-Hale, J., Saxe, R., Dungan, J., and Young, L. L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences of the United States of America*, 110(14):5648–5653.

Kourtzi, Z. and Kanwisher, N. (2001). Representation of perceived object shape by the human lateral occipital complex. *Science (New York, NY)*, 293(5534):1506–1509.

Kovera, M. B., Park, R. C., and Penrod, S. D. (1991). Jurors' perceptions of eyewitness and hearsay evidence. *Minnesota Law Review*, 76:703.

Kriegeskorte, N. and Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8):401–412.

Kryven, M., Ullman, T., Cowan, W., CogSci, J. T., and 2016 (2016). Outcome or Strategy? A Bayesian Model of Intelligence Attribution. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Lafer-Sousa, R. and Conway, B. R. (2013). Parallel, multi-stage processing of colors, faces and shapes in macaque inferior temporal cortex. *Nature Neuroscience*, 16(12):1870–1878.

Lake, B. M., Salakhutdinov, R., and Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science (New York, NY)*, 350(6266):1332–1338.

Landau, B. and Gleitman, L. R. (1985). *Language and experience: Evidence from the blind child.* Cognitive science series, 8. Harvard University Press, Cambridge, MA, US.

Lenci, A., Baroni, M., Cazzolli, G., and Marotta, G. (2013). BLIND: a set of semantic feature norms from the congenitally blind. *Behavior Research Methods*, 45(4):1218–1233.

Leopold, A., Krueger, F., dal Monte, O., Pardini, M., Pulaski, S. J., Solomon, J., and Grafman, J. (2012). Damage to the left ventromedial prefrontal cortex impacts affective theory of mind. *Social Cognitive and Affective Neuroscience*, 7(8):871–880.

Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *The Journal of neuroscience*, 31(8):2906–2915.

Lin, N., Yang, X., Li, J., Wang, S., Hua, H., Ma, Y., and Li, X. (2018). Neural correlates of three cognitive processes involved in theory of mind and discourse comprehension. *Cognitive, Affective, & Behavioral Neuroscience*, 18(2):273–283.

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., and Barrett, L. F. (2012). The brain basis of emotion: a meta-analytic review. *The Behavioral and brain sciences*, 35(3):121–143.

Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Wheelwright, S. J., Sadek, S. A., Suckling, J., MRC AIMS Consortium, and Baron-Cohen, S. (2010). Shared Neural Circuits for Mentalizing about the Self and Others. *Journal of cognitive neuroscience*, 22(7):1623–1635.

Mano, Y., Harada, T., Sugiura, M., Saito, D. N., and Sadato, N. (2009). Perspective-taking as part of narrative comprehension: A functional MRI study. *Neuropsychologia*, 47(3):813–824.

Mar, R. A. (2011). The Neural Bases of Social Cognition and Story Comprehension. *Annual review of psychology*, 62(1):103–134.

McNamee, D. and Wolpert, D. M. (2019). Internal Models in Biological Control. *Annual review of control, robotics, and autonomous systems*, 2:339–364.

Miene, P., Borgida, E., and Park, R. (1993). The evaluation of hearsay evidence: A social psychological approach. In *Individual and group decision making: Current issues.*, pages 151–166. Lawrence Erlbaum Associates, Inc, Hillsdale, NJ, US.

Mills, C. M. (2013). Knowing when to doubt: developing a critical stance when learning from others. *Developmental Psychology*, 49(3):404–418.

Mitchell, J. P., Banaji, M. R., and Macrae, C. N. (2005). General and specific contributions of the medial prefrontal cortex to knowledge about mental states. *NeuroImage*, 28(4):757–762.

Mitchell, R. E. and Karchmer, M. (2004). Chasing the Mythical Ten Percent. *Sign Language Studies*, 4(2):138–163.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., and Just, M. A. (2008). Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science (New York, NY)*, 320(5880):1191.

Moeller, M. P. and Schick, B. (2006). Relations Between Maternal Input and Theory of Mind Understanding in Deaf Children. *Child development*, 77(3):751–766.

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410.

Olson, G. (2003). Reconsidering unreliability: Fallible and untrustworthy narrators. *Narrative*, 11(1):93–109.

Ong, D. C., Soh, H., Zaki, J., and Goodman, N. D. (2019). Applying Probabilistic Programming to Affective Computing. *IEEE Transactions on Affective Computing IS - SN - VO - VL -*, pages 1–1.

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

Op de Beeck, H. P. (2010). Probing the mysterious underpinnings of multi-voxel fMRI analyses. *NeuroImage*, 50(2):567–571.

Peelen, M. V., Atkinson, A. P., and Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *Journal of Neuroscience*, 30(30):10127–10134.

Perner, J., Aichhorn, M., Kronbichler, M., Staffen, W., and Ladurner, G. (2006). Thinking of mental and other representations: The roles of left and right temporo-parietal junction. *Social Neuroscience*, 1(3-4):245–258.

Peterson, C. C., Peterson, J. L., and Webb, J. (2000). Factors influencing the development of a theory of mind in blind children. *British Journal of Developmental Psychology*, 18(3):431–447.

Peterson, C. C. and Siegal, M. (2016). Representing Inner Worlds: Theory of Mind in Autistic, Deaf, and Normal Hearing Children. *Psychological science*, 10(2):126–129.

Peterson, C. C. and Wellman, H. M. (2018). Longitudinal Theory of Mind (ToM) Development From Preschool to Adolescence With and Without ToM Delay. *Child development*, 72(1):685.

Peterson, C. C., Wellman, H. M., and Liu, D. (2005). Steps in Theory-of-Mind Development for Children With Deafness or Autism. *Child development*, 76(2):502–517.

Peterson, C. C., Wellman, H. M., and Slaughter, V. (2012). The Mind Behind the Message: Advancing Theory-of-Mind Scales for Typically Developing Children, and Those With Deafness, Autism, or Asperger Syndrome. *Child development*, 83(2):469–485.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., and Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):676–682.

Richardson, H., Koster-Hale, J., Caselli, N. K., Magid, R. W., Benedict, R., Olson, H., Pyers, J., and Saxe, R. (2019). Reduced Neural Selectivity for Mental States in Deaf Children with Delayed Exposure to Sign Language. *PsyArXiv*.

Richardson, H., Lisandrelli, G., Riobueno-Naylor, A., and Saxe, R. (2018). Development of the social brain from age three to twelve years. *Nature communications*, 9(1):1027.

Rosch, E. H. (1973). Natural categories. *Cognitive psychology*, 4(3):328–350.

Said, C. P., Moore, C. D., Engell, A. D., Todorov, A., and Haxby, J. V. (2010a). Distributed representations of dynamic facial expressions in the superior temporal sulcus. *Journal of Vision*, 10(5):11.

Said, C. P., Moore, C. D., Norman, K. A., Haxby, J. V., and Todorov, A. (2010b). Graded representations of emotional expressions in the left superior temporal sulcus. *Frontiers in systems neuroscience*, 4:6.

Saxe, R. (2005). Against simulation: the argument from error. *Trends in Cognitive Sciences*, 9(4):174–179.

Saxe, R. and Houlihan, S. D. (2017). Formalizing emotion concepts within a Bayesian model of theory of mind. *Current Opinion in Psychology*, 17:15–21.

Saxe, R. and Powell, L. J. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. *Psychological science*, 17(8):692–699.

Saxe, R. and Young, L. (2013). Theory of Mind: How brains think about thoughts. In Ochsner, K. and Kosslyn, S., editors, *The Handbook of Cognitive Neuroscience*, pages 204–213. Oxford University Press.

Saxe, R. R. and Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *NeuroImage*, 19(4):1835–1842.

Saxe, R. R., Whitfield-Gabrieli, S., Scholz, J., and Pelphrey, K. A. (2009). Brain regions for perceiving and reasoning about other people in school-aged children. *Child development*, 80(4):1197–1209.

Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., and Adolphs, R. (2015). Deconstructing and reconstructing theory of mind. *Trends in Cognitive Sciences*, 19(2):65–72.

Schick, B., de Villiers, P., de Villiers, J., and Hoffmeister, R. (2007). Language and Theory of Mind: A Study of Deaf Children. *Child development*, 78(2):376–396.

Schick, B. and Hoffmeister, R. (2001). ASL skills in deaf children of deaf parents and of hearing parents. In *Society for Research in Child Development International Conference, Minneapolis, MN*.

Scholl, B. J. and Leslie, A. M. (2001). Minds, Modules, and Meta-Analysis. *Child development*, 72(3):696–701.

Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E. N., and Saxe, R. (2009). Distinct Regions of Right Temporo-Parietal Junction Are Selective for Theory of Mind and Exogenous Attention. *PLoS ONE*, 4(3):e4869–7.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., and Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, 42:9–34.

Sebastian, C. L., Fontaine, N. M. G., Bird, G., Blakemore, S.-J., De Brito, S. A., McCrory, E. J. P., and Viding, E. (2012). Neural processing associated with cognitive and affective theory of mind in adolescents and adults. *Social Cognitive and Affective Neuroscience*, 7(1):53–63.

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

Shamay-Tsoory, S. G. (2011). The Neural Bases for Empathy. *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry*, 17(1):18–24.

Shamay-Tsoory, S. G. and Aharon-Peretz, J. (2007). Dissociable prefrontal networks for cognitive and affective theory of mind: a lesion study. *Neuropsychologia*, 45(13):3054–3067.

Shamay-Tsoory, S. G., Tibi-Elhanany, Y., and Aharon-Peretz, J. (2006). The ventromedial prefrontal cortex is involved in understanding affective but not cognitive theory of mind stories. *Social Neuroscience*, 1(3-4):149–166.

Shum, M., Kleiman-Weiner, M., Littman, M. L., and Tenenbaum, J. B. (2019). Theory of Minds: Understanding Behavior in Groups Through Inverse Planning. *arXiv.org*.

Siegal, M. and Peterson, C. C. (1994). Children's theory of mind and the conversational territory of cognitive development. *Children's early understanding of mind: Origins and development*, pages 427–455.

Skerry, A. E. and Saxe, R. (2014). A common neural code for perceived and inferred emotion. *The Journal of neuroscience*, 34(48):15997–16008.

Skerry, A. E. and Saxe, R. (2015). Neural Representations of Emotion Are Organized around Abstract Event Features. *Current Biology*, 25(15):1945–1954.

Sommer, M., Döhnel, K., Sodian, B., Meinhardt, J., Thoermer, C., and Hajak, G. (2007). Neural correlates of true and false belief reasoning. *NeuroImage*, 35(3):1378–1384.

Spotorno, N., Koun, E., Prado, J., Van Der Henst, J.-B., and Noveck, I. A. (2012). Neural evidence that utterance-processing entails mentalizing: The case of irony. *NeuroImage*, 63(1):25–39.

Spreng, R. N., Mar, R. A., and Kim, A. S. N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. *Journal of cognitive neuroscience*, 21(3):489–510.

Spunt, R. P., Kemmerer, D., and Adolphs, R. (2015). The neural basis of conceptualizing the same action at different levels of abstraction. *Social Cognitive and Affective Neuroscience*, 11(7):1141–1151.

Tamir, D. I. and Thornton, M. A. (2018). Modeling the Predictive Social Mind. *Trends in Cognitive Sciences*, 22(3):201–212.

Tamir, D. I., Thornton, M. A., Contreras, J. M., and Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences of the United States of America*, 113(1):194–199.

Tanaka, K. (1993). Neuronal mechanisms of object recognition. *Science (New York, NY)*, 262(5134):685–688.

Thornton, M. A. and Tamir, D. I. (2017). Mental models accurately predict emotion transitions. *Proceedings of the National Academy of Sciences of the United States of America*, 114(23):5982–5987.

Ullman, T. D., Baker, C. L., Macindoe, O., Evans, O., Goodman, N. D., and Tenenbaum, J. B. (2009). Help or Hinder: Bayesian Models of Social Goal Inference. In Bengio, Y., Schuurmans, D., Lafferty, J. D., Williams, C. K. I., and Culotta, A., editors, *Advances inNeural Information Processing Systems*, pages 1874–1882. Curran Associates, Inc.

van Ackeren, M. J., Casasanto, D., Bekkering, H., Hagoort, P., and Rueschemeyer, S.-A. (2012). Pragmatics in Action: Indirect Requests Engage Theory of Mind Areas and the Cortical Motor Network. *Journal of cognitive neuroscience*, 24(11):2237–2247.

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., Maier, W., Shah, N. J., Fink, G. R., and Zilles, K. (2001). Mind Reading: Neural Mechanisms of Theory of Mind and Self-Perspective. *NeuroImage*, 14(1):170–181.

Wen, T., Mitchell, D. J., and Duncan, J. (2019). The functional convergence and heterogeneity of social, episodic, and self-referential thought in the default mode network. *bioRxiv*, 1:753509.

Wimmer, H. and Perner, J. (1983). Beliefs about beliefs: representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, 13(1):103–128.

Winecoff, A., Clithero, J. A., Carter, R. M., Bergman, S. R., Wang, L., and Huettel, S. A. (2013). Ventromedial Prefrontal Cortex Encodes Emotional Value. *Journal of Neuroscience*, 33(27):11032–11039.

Woolfe, T., Want, S. C., and Siegal, M. (2002). Signposts to development: Theory of mind in deaf children. *Child development*, 73(3):768–778.

Wu, Y., Baker, C. L., Tenenbaum, J. B., and Schulz, L. E. (2018). Rational Inference of Beliefs and Desires From Emotional Expressions. *Cognitive Science*, 42(3):850–884.

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., Roffman, J. L., Smoller, J. W., Zöllei, L., Polimeni, J. R., Fischl, B., Liu, H., and Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, 106(3):1125–1165.

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