Cortex

Review Paper

Building Blocks of Social Cognition:

Mirror, Mentalize, & Share?

Daniel Alcalá-López[1,2], Kai Vogeley[3,4], Ferdinand Binkofski[5], Danilo Bzdok[1,2,6]

[1] Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen University, Germany
[2] JARA, Translational Brain Medicine, Aachen, Germany
[3] University Hospital Cologne, Department of Psychiatry and Psychotherapy, Germany
[4] Institute of Neurosciences and Medicine – Cognitive Neuroscience (INM3), Research Center Juelich, Germany
[5] Division for Clinical and Cognitive Neurosciences, Department of Neurology, RWTH Aachen University, Germany
[6] Parietal team, INRIA, Neurospin, bat 145, CEA Saclay, 91191 Gif-sur-Yvette, France

Keywords: empathy; theory of mind; mirror neuron system;

Abstract

During the past decade, novel approaches to study social interaction have expanded and questioned long-standing knowledge about how humans understand others. We aim to portray and reconcile the key psychological processes and neural mechanisms underlying navigation of the social environment. Theoretical accounts mostly revolved around either abstract inferences or embodied simulations, whereas experimental studies mostly focused on theory of mind, empathy, and action imitation. The incongruences between theories of and experiments on social cognition are revisited to reduce uncertainty. We finally retrace differential impairments in social capacities as a means to re-conceptualize psychopathological disturbance in psychiatry, including schizophrenia, borderline personality, and autism.

Introduction

During recent years, our understanding of the cognitive processes and neurobiological mechanisms underlying human capacities in social dynamics have been continuously challenged and diversified. For instance, many investigators believed the ability to understand the motivation behind others' behavior to emerge around the age of 4. More recent evidence suggests its presence in infants as young as 2 years (for a review, see Scott and Baillargeon 2017), and certain nonhuman primates could make use of comparable abilities (Krupenye and others 2016). Additionally, the use of the umbrella term "empathy" denoting multiple affective, cognitive, and motivational processes is increasingly criticized as an impediment to scientific progress and communication (Bloom 2017).

We juxtapose the major theoretical accounts of how individuals understand the emotions, beliefs, and intentions of others, which mostly concentrated on either making abstract inferences or automatic simulations of others' mental states. We then examine the key experimental approaches to the study of social interaction, which concentrated on studies of theory of mind, empathy, and action imitation. The present work aims at clarifying the precise mapping between these *two theoretical accounts* and *three experimental research streams* that remains a conundrum in the social and affective sciences.

Starting points in social neuroscience: Two process models of human social cognition

Successful social interaction requires adequate exploration and modeling of other individuals' mental or "inner" states, including their emotions, beliefs, or intentions behind action (Frith and Frith 1999; Mitchell 2009). Two main theoretical frameworks have emerged to explain how an adequate understanding of others' behavior may be achieved. On the one hand, the *theory theory* (TT) proposes that individuals draw on a collection of abstract principles about human behavior, acquired through life experience, that allows to interpret and predict the mental states and behavioral patterns of others (Carruthers 2009; Gopnik and Wellman

1994). Such inference based on *abstract emulation* enables predictions about what individuals are thinking about others, despite the fact that their mental states can never be directly observed or explicitly confirmed (Gopnik and Wellman 1992; Perner 1991; Wellman 1990). As defined by Saxe (2005), the TT account captures humans' lay theory of psychology constructed "from observation, inference and instruction, and then deployed to predict or explain another person's inference, decision or action".

On the other hand, the *simulation theory* (ST) proposes that humans partially impersonate others and automatically imitate their mental states (Gallese and Goldman 1998; Goldman 2000). This neurocognitive reinstantiation of the other's mental content aims to understand what oneself would experience in the other's place (Gallagher 2001). According to the ST account based on *embodied simulation* **(Gallese and Sinigaglia 2011)**, grasping the mental states of others means to "purposely pretend to be in the other's 'mental shoes' and use our own mind as a model for the mind of others" (Gallese 2003). As such, the ST framework rejects the need to assemble abstract models to emulate others' behavior –as proposed by the TT– since humans already have a working model of how it feels to perceive and act in a given environment: one's own inner experience.

In the following, previous behavioral and neuroscience research inspired by TT and ST will be portrayed, as well as more recent theoretical accounts (e.g. Gallagher and Hutto 2008). We will then discuss both conceptual frameworks with respect to their neural mechanisms.

Theory theory: Conceptual emulation of others' mental states

The TT received extensive empirical support from experimental studies in developmental and comparative psychology focusing on perspective-taking tasks that require deception and false-belief detection (Baron-Cohen and others 1985; Leslie 1987; 1994; Premack and Woodruff 1978; Wimmer and Perner 1983). For instance, Premack and Woodruff (1978) originally referred to the awareness that human and non-human primates may have of others' mental states as *theory of mind* (ToM). Frith and colleagues (1991) later introduced the term *mentalizing* to refer to the ability of belief attribution in humans and to include spontaneous and non-inferential capacities, as the term "theory" could lead to the misunderstanding that ToM is a purely theoretical or intellectual account.

Whether non-human primates can adopt, at least to some extent, the perspective of other conspecifics has been debated at length in the past decades (Call and Tomasello 2008; Premack and Woodruff 1978). Tomasello and colleagues (2003) described how previous research during the 1990s led many investigators to believe that such a capacity would be a defining feature of the human species. However, more recent findings since the beginning of the century have challenged this view, and they rather entailed a continuous refinement of the comparative difference in ToM capacities in monkeys and humans. For instance, Hare and colleagues designed a series of experiments in which two chimpanzees, one dominant and one subordinate, had to compete for food. By strategically hiding the food in locations to which only one or both chimpanzees had visual access, the authors investigated whether the subordinate was aware of what the dominant could or could not perceive at the moment (Hare and others 2000) or in past situations (Hare and others 2001). This is in line with a recent behavioral study in three different species of apes (Krupenye and others 2016) that used an anticipatory looking measure as a rudimentary proxy to test for false-belief understanding. The behavioral experiments showed that apes not only could infer the goals and intentions of others' (external) actions, but also behaved in alignment with their (internal) mental states that were incongruent with the external reality (i.e., false beliefs). It is therefore becoming increasingly difficult to precisely demarcate the difference between mind reading capacities in humans and other primate species.

Many experimental psychology studies in humans exemplified the capacity to attribute intentions and beliefs to others (i.e., TT) by using experimental paradigms based on the concept of ToM, such as false beliefs tasks (Baron-Cohen and others 1985), advocated by many as a fundamental mechanism underlying social interaction (Carruthers and Smith 1996; Frith and Frith 2003). Traditional experimental paradigms prompting participants to indicate the beliefs of others found that ToM emerges around the age of 4 years (Wellman and others 2001). This capacity was thus regarded as a relatively advanced form of social cognition. However, in the past decade novel approaches have shifted the focus from probing how children answer specific test questions about their reaction to experimental situations (Baillargeon and others 2016). For instance, Buttelmann and others (2009) taught in a classical false-belief task a group of infants younger than 2 years of age how to unlock a pair of boxes with a pin. Afterwards, one of two experimenters would put a toy in one of the boxes and leave the room with the box unlocked. The second experimenter would then put the toy into the other box and lock both boxes. The first experimenter on his return would try to get the toy from the original box in which he had previously put the toy, while the infants were then prompted to help the experimenter. The authors found that the infants succeeded in helping to retrieve the toy by unlocking the box where the toy had been hidden, indicating awareness of the experimenter's false belief. Based on similar experiments, several recent studies found that children younger than 2 years of age already behaving according to others' false beliefs (for a review see: Scott and Baillargeon 2017), though the developmental trajectory of which is probably culture-dependent (Vogeley and Roepstorff 2009).

On the other side of the lifespan, there has been increasing interest in the study of age-related differences in ToM abilities. Despite initial findings of a better performance in ToM tasks in a group of elderly compared with younger participants (Happé and others 1998), later studies have found contradicting results. For instance, Maylor and others (2002) reported an age-related decline in ToM abilities. This is congruent with a recent meta-analysis of 23 studies that included six different types of ToM tasks (Henry and others 2013) that showed a moderate (*r* = –.41) deficit in performance across ToM tasks and stimuli presentation modality in older adults.

Furthermore, such age-related deficit was larger in magnitude than corresponding deficits in matched control tasks, which led the authors to suggest that ToM may be a domain-specific process, which declines with age disregarding perceptual or linguistic capacities.

Taking the behavioral findings to the neural level, the brain-imaging literature on the TT framework has also frequently relied on ToM tasks. Such neuroimaging studies have consistently revealed that a set of brain regions (Fig. 1) including the medial prefrontal (mPFC) and posterior cingulate (PCC) cortices, as well as the bilateral temporo-parietal junction (TPJ), robustly increase in neural activity when participants undergo perspective-taking tasks probing ToM performance (Gallagher and Frith 2003; Saxe and others 2004; Saxe and Kanwisher 2003; Vogeley and others 2001). A virtually identical set of brain regions is, however, also known to increase its activity during the retrieval of autobiographical memory, spatial navigation, or prospection into the future (Buckner and Carroll 2007; Spreng and others 2009; Vogeley and Fink 2003; Vogeley and others 2004). In spite of their involvement in diverse experimental conditions, this set of regions was previously found to decrease in neural activity during many other tasks and was therefore called the default mode network (DMN), mostly active during idling mind sets (Gusnard and Raichle 2001; Raichle and others 2001). **This considerable overlap between the neural correlates of the resting baseline and those of ToM as a fundamental kind of social cognition has led some authors to suggest that social information processing may frequently be part of the physiological resting state of the brain (Schilbach and others 2008; Vogeley 2017).**

Simulation theory: Embodied imitation of others' mental states

Despite extensive empirical support in favor of TT, some authors have denied that a mechanism dedicated to abstract emulation is a necessary condition to grasp and represent others' subjective experience (Perner and Kühberger 2005). Instead, the ST account proposes that individuals automatically mimic or intuitively impersonate in a covert fashion the behavior of others, even when simply observing them (Fogassi and Ferrari 2007; Umilta and others 2001). ST proposes that this reinstantiation of observed behavior enables access to the internal mental state of the other, thus enabling action understanding (Gallese and others 2004; Keysers and Gazzola 2007; Uddin and others 2007). The ST framework for grasping other humans' minds has often been used as a conceptual basis to interpret experimental studies on empathy tasks. Consistently, Preston and De Waal (2002) proposed that witnessing others' social-affective behavior inevitably triggers one's own internal representation of that same behavior. Most researchers probably agree on a working definition of empathy as consciously experiencing an affective mental state that is congruent or very comparable to that of an observed individual (De Vignemont and Singer 2006; Decety and Chaminade 2003).

Developmentally, simpler forms of affective sharing were suggested to precede the onset of full-fledged empathy capacities in infants (Singer and Lamm 2009). Concretely, mimicry and emotional contagion are already present in newborns (Piaget 1945), before the onset of ToM (for a review, see Meltzoff and Moore 1989). Dimberg and Öhman (1996) for instance investigated facial expressions using electrophysiological measures to show triggering of corresponding facial gestures (e.g., smiling or frowning) when perceiving others' affective expressions. Such a tendency to automatically reproduce the externally visible manifestations of internal affective states (i.e., mimicry) has been suggested as a possible low-level mechanism, elaborated on by more complex forms of empathy (Hatfield and others 1993; Singer and Lamm 2009). In emotional contagion, another proto-form of empathy, an individual synchronizes with and converges to others' affective mental states (Hatfield and others 2009). In contrast to full-blown empathy, emotional contagion occurs without awareness of the observing individual (de Waal 1999; Decety and Jackson 2004).

On the neural level, empathic state-matching reaction to others' affective behavior has consistently been associated with a brain network including the anterior insula (AI) and anterior mid-cingulate cortex (aMCC). This so-called saliency network is recruited, for instance, both when a participant receives painful stimulation as well as when perceiving others in pain (Decety 2010; Fan and others 2011; Lamm and others 2011; Singer and others 2004). A majority of studies in the neuroimaging literature on empathy tasks performed such comparison between the neural activity elicited by the observation of others in pain and by experiencing pain oneself (Decety and Lamm 2006; Singer and Lamm 2009; Singer and Leiberg 2009). This concurs with the idea that social cues can elicit partial synchronization of neural activity patterns in agent and observer (Adolphs 2003; Decety and Grèzes 1999; Gallese 2003).

Mirroring Neural Spikes of Others' Behavior

We argue for a broader notion of ST going beyond affective sharing. Automatic simulation of *affective* mental states according to the ST framework has largely focused on empathy and mechanisms of sharing emotion in the social neuroscience literature. However, *affect-* and *emotion-independent* mechanisms for simple action observation can be readily viewed as another flavor of internally reenacting others' behavior. **This is not necessarily an argument against a simulation-based account for understanding others, but a notion compatible with the above-mentioned original description of the ST**. This is suggested by invasive experimental findings in monkeys that showed existence of neuron populations, called mirror neurons, that fire in response to both executing and observing the same goal-directed action (Di Pellegrino and others 1992; Fogassi and others 1998; Gallese and others 1996).

By *in-vivo* recordings of single-neuron activity in the ventral premotor cortex of macaque monkeys, Rizzolatti and others (1996) found a subset of neurons that discharged when the monkeys grasped, held, or placed an object, as well as when it was the experimenter who was performing such actions. Consistently, Kohler and others (2002) found in recordings in macaques that this matched firing pattern was not exclusively evoked by visual, but also by auditory stimuli. Further, mirror neurons seem to discharge exclusively during goal-directed hand actions (e.g., grasping, tearing, and holding) but not in response to goal-free muscle contractions (Gallese and others 1996). In these experiments unrelated to mimicry, emotion contagion, or empathy, only observing a goal-directed action triggers the neural activity pattern responsible for execution of that same action in the observer's brain. These empirical findings in monkeys have frequently enticed speculation that humans understand the actions of conspecifics because a human mirror-neuron analog estimates possible outcomes of observed actions (Gallese and others 2004).

In humans, neuroimaging techniques allowed for noninvasive exploration whether identical brain regions are recruited during passive perception and active execution of particular actions (Buccino and others 2001; Iacoboni and others 1999; Nishitani and Hari 2000). Direct evidence for mirror neurons in humans is, however, seldom due to ethical constraints around electrophysiological recordings in healthy participants (Mukamel and others 2010). Nevertheless, it was suggested that a putative “mirror neuron system” (MNS) in humans could contribute to understanding others’ actions and their underlying causes by internal neural simulation (Gallese and Goldman 1998; Gallese and others 2004; Rizzolatti and others 2001). Etzel and colleagues (2008) provided fMRI evidence for similar neural activity patterns in the premotor cortex during the execution and perception of an action. The authors used multivariate learning algorithms to classify the neural activity pattern of the premotor cortex when participants had to discriminate between the sound of a hand or mouth action in a similar task to that described by Gazzola and colleagues (2006). Once trained, the classification algorithm could determine whether the participant was executing a hand or mouth action at a later moment of the experiment. In sum, electrophysiological and neuroimaging experiments emphasize the link between the perceived actions of others and their automatic reproduction in the observer, which invigorates the idea that the MNS allows understanding others by subliminal or subconscious re-experiencing or re-instantiating their behavior, compatible with the ST account.

Although the MNS was originally suggested to account for simulation of motor actions in single-cell recording experiments in monkeys (Chersi and others 2011; Fogassi and others 2005), evidenced neural simulations of emotion-unrelated motor action have often been extended to also explain a variety of social-affective psychological phenomena such as ToM and empathy (Goldman 1992; 2006; Gordon 1986). Indeed, several authors have recently extended the MNS-based account of neural simulation to include empathic processes (Gallese and others 2004; Keysers and Gazzola 2009; Pfeifer and others 2008) due to previous findings showing that empathy for pain involves at least some of the components of pain perception in an fMRI study (Singer and others 2004). Additionally, the

ToM system again comes into play as soon as perceived movements can no longer be interpreted on the basis of expectations, but are different from what was anticipated (Georgescu and others 2014). This can be taken to suggest that the TT-related perspective-taking system can be recruited to supplement the MNS and other instances of ST-related processes like empathy –two closely related and intertwined systems that are not mutually exclusive.

Even though parsimonious in principle, the ST has received numerous critics and revisions (Brass and others 2007; Jacob and Jeannerod 2005; Kilner 2011; Newen and Schlicht 2009). For instance, Mitchell and colleagues (2006) pointed out that a simulation mechanism would be necessarily limited to real-time social interactions during which one can perceive the other´s current physical states. However, those mental states that derive from previous knowledge of attitudes or long-term dispositions cannot be inferred from observed external behavior (Mitchell and others 2006). Similarly, certain social behaviors may be less dependent on inferences from real-time sensory input. Umilta and colleagues (2001) reported that half of the mirror neurons they recorded in the premotor cortex in monkeys would fire not only in response to action observation, but also when the final part of the movement was blocked. The authors speculated that this subpopulation of mirror neurons would represent actions even if no actual movements are perceived in the environment. However, this animal experiment was not performed during the total absence of sensory input, but only during a limited time window. Therefore, the extent to which the MNS can simulate others’ behavior without sensory input of others' ongoing motor action currently remains unclear.

Moreover, authors supporting a TT view have argued that ST cannot account for the systematic errors children make when attributing mental states to others (Nichols and Stich 2003; Saxe 2005). For instance, Ruffman (1996) showed in a behavioral study a set of beads grouped by color in different bowls to a group of children. He put a bead in a box, and then asked the children to guess what color would another individual think the bead was if he or she could not see from which bowl he took the bead. The results showed that children being 4 years old or younger would more often erroneously ascribe false beliefs to other individuals.

Such finding suggests that children develop naïve rules about others' beliefs, such as "perceiving entails knowing".

Congruently, Ramnani and Miall (2004) found evoked neural activity in the dorsal premotor cortex (PMd) during the preparation of responses in a Pavlovian associative task in which arbitrary visual cues determined future actions. In contrast, activity related to the anticipation of the responses of another individual did not activate the PMd, but the dorsal mPFC, TPJ and connected parts of the premotor cortex instead. Unlike other studies, Ramnani and Miall (2004) did not ask participants to explicitly attribute mental states to others. Rather, they provided explicit, simple rules that participants had to learn and their application was evaluated during the experimental condition. Thus, the authors ensured that participants could certainly anticipate the actions of other individuals. Despite the simple nature of the task, the activity in the dorsal mPFC and TPJ is consistent with a TT-related mechanism to anticipate others' behavior given that these regions have been consistently associated with ToM in perspective-taking tasks (Gallagher and Frith 2003; Saxe and others 2004; Saxe and Kanwisher 2003).

Therefore, depending on the nature of the task, the ST may be inadequate to understand and predict the actions of another individual. A possible explanation might be that the ST is a necessary but not sufficient mechanism to understand others in situations. This becomes especially clear in situations in which we can no longer predict or anticipate the outcome. Unexpected outcomes that do not match our assumptions plausibly require TT, rather than ST, mechanisms.

Integrative concepts to explain social behavior

The TT and ST frameworks were introduced and treated as conceptual and empirical opponents based on evidence from developmental psychology, functional neuroimaging, and theoretical reasoning. This led many investigators in experimental psychology and neuroscience to take sides with either the ST or TT position (Carruthers 1996; Goldman 1992). However, behavioral and neural findings have encountered difficulties in settling whether the ST or TT is the predominant mechanism for explaining and predicting other individuals' behavior and their corresponding brain manifestations. In an fMRI study by Grèzes and colleagues (2004), neural activity latency in the PMd and TPJ was higher when perceiving the actions of others compared to those of oneself. This is congruent with the ST account given that the neural mechanism involved in perceiving and simulating an action is the same when the observer is also the agent. However, the authors also found activity in ToM-related regions when the observer inferred the agent's false beliefs about motor action, more coherent with TT than ST. Collectively, these results cannot unequivocally support either ST or TT as a unique explanation of understanding others' and one's own actions, suggesting that humans make combined use of both systems in everyday-life social dynamics.

The question whether the ability to attribute mental states to others is implemented by simulating (i.e. ST) or making inferences about (i.e. TT) others' behavior raised renovated interest in the last years due to emerging computational approaches. Since there is a large tradition of computational approaches to study reward learning and decision making processes, many authors have designed experimental paradigms in which participants were required to learn the contingencies of rewards from the observation of others' behavior. For example, Behrens and others (2009) presented a learning game in which participants had to ascertain the likelihood of a reward's location as well as the reliability of a partner's advice regarding said location. The authors found that a simple associative learning model could explain how participants updated their beliefs on both the location and the reliability of their partner through prediction errors. Moreover, the authors found that social prediction error signals correlated with neural activity patterns within the pSTS/TPJ, a brain region considered as part of the ToM-related network (Saxe and Kanwisher 2003; Van Overwalle 2009). Nevertheless, others have switched the focus towards the study of predictions about goal-directed actions, to other people's beliefs (i.e. ToM), or even to personality traits (for a review, see Koster-Hale and Saxe 2013). For instance, Hampton and others (2008) designed a strategic game to investigate whether different computational models explained the neural activity patterns of ToM-related regions. During the game, participants

alternated between the role of an employer or an employee. While the former could decide whether to inspect the employee or not, the latter could either work or avoid working at the risk of getting caught. These authors found that neural activation in the mPFC at the time of choice correlated with the prediction each participant made about their opponent's intentions. Furthermore, activation in the STS/TPJ at the time of the outcome correlated with the deviation of each participants' behavior from the prediction that their opponent had made (i.e. the prediction error). These results support the notion of ToM as an inferenced-based mechanism to understand the intentions of others in line with TT. In a similar fashion, Yoshida and others (2008) used a multi-player game where a human participant could either cooperate with the opponent (a computer agent) to hunt a large prey, or hunt independently a small prey. In this 'stag-hunt' game, players need to infer the other's goals to optimize their own behavior and, in doing so, they must consider the inferences the opponent makes for their own strategy as well. That is, player A (human participant) needs to adjust his or her choice behavior by considering the inferences that player B (computer agent) makes regarding the strategy of player A, and vice versa. The model introduced by the authors was designed to estimate the depth of the inferences that players make about the computer agent's beliefs. The results showed that such model of choice behavior outperformed others that did not account for any inference, congruently with the TT framework. The importance of this recursive nature of the inferences we make to understand others' intentions have been further supported in more recent studies (de Weerd and others 2015; Devaine and others 2014). Overall, computational models of mental state attribution largely coincide in showing the likelihood of an inference-based mechanism when participants engage in highly demanding tasks. As defined by Baker and others (2011), "ToM inferences come surprisingly close to those of an ideal rational model, performing Bayesian inference over beliefs and desires simultaneously".

Besides the now classical ST and TT accounts, many alternative, often integrative theoretical accounts have emerged over the past decade to accommodate the shortcomings of the two classic views. Gallagher (2008) proposed a perceptual mechanism based on the premise that others' external behavior is a direct expression of their mental states. The so-called *direct perception theory* implies that when perceiving socially relevant environmental information (e.g., faces or body movements), a pattern-matching mechanism dependent on previously learned stimuli configurations would detect the behavioral fingerprint linked to a specific mental state. As such, neither abstract inference nor bodily simulation would be required to understand other individuals' intentions, in contrast to the TT and ST accounts.

Another alternative account for social cognition, referred to as *narrative practice hypothesis* (Hutto 2008), accredits a fundamental role in the understanding of another's behavior to how humans acquire knowledge about them through narrative stories. In a later redefinition, both authors combined their previous acquaintances to state that not only individuals understand from direct observation of other individuals' behavior (as originally suggested by Gallagher), but also get involved in interactive situations with them. From these interactions, humans learn about others by narratives which, in turn, provide abstract background knowledge for future social interactions (Gallagher and Hutto 2008). Overall, a common feature of recent accounts for social interaction is that individuals do not appear to depend on a single mechanism to understand others, but on at least two or more complementary mechanisms. It is their elusive nature what remains a topic of debate: low- versus high-level simulation (Goldman 2006), intuition versus inference-based understanding (Gallagher 2001), or implicit versus explicit modeling (Newen 2014).

There is hence lack of solid experimental or neuroscientific evidence that could tip the balance between different candidate mechanisms to explain social behavior. Some authors are therefore calling into question the long-assumed opposition between TT and ST by pointing out that understanding others might involve two different, yet synergistic and complementary mechanisms that simply have two different functional roles in interacting and communicating with others (Apperly 2008; De Lange and others 2008; Kilner 2011; Van Overwalle and Baetens 2009). Brass and colleagues (2007) concluded from a functional brain-imaging study that a MNS-related mechanism did not mediate action understanding when the observed action was novel or when it was hard to understand. Instead, as stated before, a mental-state inference mechanism may be required. The authors further argued that it is the contextual plausibility that determines whether the observer can map the target's behavior based on own motor schemes in stereotypical, highly familiar actions, or they would need to explicitly infer the purpose of an unusual action in novel contexts. In another neuroimaging study Santos and colleagues (2010) showed that gradual induction of a sense of animacy (via biological movement) recruited different key regions of the social brain. Abstract, inert objects were perceived as animated only on the basis of changing motion parameters that were highly suggestive of personal agents "behind" the movements governing them. While the evaluation of actually present animacy signals recruited the vmPFC, AI, STS, FG, and HC, the mere disposition to detect socially salient movements were associated with increased neural activity in the superior parietal lobe and ventral PM, thus resembling MNS. This might point to a putative gradient of complexity between TT-associated mechanisms of abstract emulation and ST-associated mechanisms of embodied simulation underlying social interaction. Thus, while understanding simple motor actions performed by others might involve a specific, lower-level neural mechanism, the more hidden the intention behind the motor act becomes, the more a higher-level neural mechanism would be needed (Vogeley, in press).

Concluding remarks

Investigations of how humans understand each other aim to reveal the natural kind or natural kinds underpinning social cognition. Theoretical accounts traditionally focused on two candidate explanations: individuals either rely on inferences (*theory theory*) or embodied simulations (*simulation theory*). More recently, several integrative theoretical accounts combining features of both mechanisms have been proposed. Congruent with this, we have argued that many experiments on ToM were mostly characterizing TT-linked mechanisms, whereas experiments on empathy versus action imitation characterized emotion-dependent versus emotion-independent mechanisms linked to ST. The precise mapping between the two theoretical accounts and the three experimental findings has long remained a conundrum in the social and systems neuroscience communities. However, widespread agreement on these distinctions is essential not only to understand the human condition, but also to decipher specific patterns of failed social interactions in major psychiatry disorders. Future research should insist on directly comparing TT and ST in brain studies. This research agenda will clarify how these candidate mechanisms to understanding the mental states of others are differentially implemented at the neural level as well as to delineate disorder-specific endo-phenotypes that could benefit medical care. As first steps, it will thus be crucial that neuroscientists and psychiatrists, as well as comparative, developmental, and experimental psychologists should adopt a commonly shared language hygiene for terms and concepts.

Figures

Figure 1: Key functional networks of the social brain. Depicts topographical overlap between meta-analytical maps of brain regions previously reported in the literature on empathy (*red*), theory of mind (ToM; *blue*), and the mirror neurons system (MNS; *yellow*). The three functional network maps are displayed separately on sagittal, coronal, and axial views of a T1-weighted MNI template rendered using Mango (multi-image analysis GUI; http://ric.uthscsa.edu/mango/). Crosshairs in the axial and coronal maps mark the location of converging areas highlighted in the sagittal maps. Publicly available data from Bzdok and others (2012) and Caspers and others (2010) can be obtained for visualization and reuse from the data-sharing platform ANIMA (http://anima.fz-juelich.de/).

Box 1: Does affective theory of mind equate with cognitive empathy?

The relationship between the TT and ST is not self-evident. Both conceptual frameworks, as well as corresponding psychological notions of ToM and empathy, have been previously used interchangeably (Baron-Cohen and others 2001; Gillberg 1992; Roeyers and others 2001). Furthermore, many authors have proposed a distinction between the cognitive versus affective aspects in ToM (Choi-Kain and Gunderson 2008; Shamay-Tsoory and Aharon-Peretz 2007) and empathy (Harari and others 2010; Zaki and Ochsner 2012). This has led to increasing confusion about the distinct core mechanisms that underlie human social behavior.

On the one hand, the TT framework implies abstract rules about how mental states explain our own and others' behavior. Although this theoretical account has largely been characterized by experimental studies of *ToM* abilities, many authors use the term *cognitive empathy* to refer to the ability to understand –yet not share– the affective state of others (e.g. Dziobek and others 2008). However, previous research suggests that the concepts of ToM and cognitive empathy are intimately related on the brain level (De Waal 2008). Patients with lesions in the mPFC are often reported to perform poorly in tasks tapping on ToM and on cognitive empathy (Eslinger 1998; Shamay-Tsoory and others 2003). Based on differential lesions of the ventral mPFC and the inferior frontal gyrus, Shamay-Tsoory and colleagues (2009) found a behavioral-anatomical double dissociation between the cognitive and affective aspects of empathy, respectively. This is in line with a recent meta-analysis showing that while empathy for pain tasks congruently involved an increased activity of the aMCC and AI, many experimental paradigms recruited other brain networks as well (Lamm and others 2011). For instance, when asked to empathize with the affective state of others based on abstract visual cues, individuals consistently engaged areas associated with ToM, including the ventral mPFC and TPJ). Whether ToM and cognitive empathy refer to the same underlying capacity or represent separate, but closely interrelated capacities to infer the internal state of others (affective or not) remains a topic for future research.

On the other hand, the notion of empathy is often applied to contexts that go beyond sharing of affective states in the here and now towards understanding the past and future of the social behavior of others, such as in tasks that require knowing others' internal states (Eslinger 1998; Zahn-Waxler and others 1992) or imagining what others are feeling or thinking (Adolphs 1999; Ruby and Decety 2004). The lack of an agreed-upon and systematically used definition of empathy in the neuroscientific literature has often led to the inclusion of a variety of concepts, experimental paradigms, and psychological facets that range from mutually exclusive to contradictory. For instance, Zaki and Ochsner (2012) understand empathy as an overarching term that involves affective 'experience sharing' (equivalent to the notion of empathy described above), but also ToM and prosocial concern (i.e., a motivation towards improving others' well-being). Indeed, Bloom (2016) recently discussed that the term 'empathy' be useful as long as the underlying processes are clearly defined and, consistently employed, while other authors argued for the suitability of empathy in the form of an umbrella term that subsumes emotional sharing, empathic concern, and affective perspective-taking (Decety and Cowell 2014).

As one possibility, psychological processes underlying TT versus ST as well as the corresponding aspects of ToM versus empathy can be conjointly recruited and activated in many real-world social interactions (Keysers and Gazzola 2007). This view is for instance supported by an fMRI study showing that imagining a loved-one compared with a stranger in painful situations produces greater activity in the often-empathy-related saliency network and less activity in regions important for self/other distinction (Cheng and others 2010), in close relation to ST. However, the authors also found that the often ToM-related regions including mPFC, TPJ, and superior frontal gyrus significantly increase activity when imagining a stranger in pain, in closer relation to TT. Thus, contextual demands can differentially involve complementary mechanisms of social cognition with relation to either TT or ST. This concurs with a recent study by Kanske and others (2016) showing the capacities to empathize with and take the perspective of others are independent, both on a behavioral and neural level.

Box 2: TT versus ST impairment in schizophrenia

During the past decade, an increasing number of behavioral and neuroimaging studies have investigated the social cognition deficits in schizophrenia patients. Such psychiatric patients often report similar affect sharing abilities than healthy participants (Achim and others 2011; Michaels and others 2014). This indicates a preserved capacity to internally simulate the affective state of others, a capacity falling under the ST framework. Furthermore, Michaels and colleagues (2014) found that some schizophrenia patients report to be especially sensitive and reactive to the affective state of others compared with healthy controls. In contrast, different meta-analyses of behavioral studies show a clear deficit in these patients' ability to understand the intentions and beliefs of others (Bora and others 2009; Savla and others 2012; Sprong and others 2007). Several behavioral findings therefore suggest that schizophrenia patients are relatively more impaired in their ability to interpret and predict the mental states of others related to TT as opposed to ST.

This suspicion is reinforced by a number of neuroimaging studies. In a recent fMRI experiment, Horan and colleagues (2016) presented videos of people in pain to a group of schizophrenia patients. They found similar neural activity of the aMCC and AI in both patients and healthy controls, supporting the idea of a preserved capacity to match the affective state of others with their own internal state. However, the authors also administered a 'self-other' condition where healthy controls displayed an increased activity in these regions during the 'self' compared with the 'other' condition, while schizophrenia patients showed the opposite pattern: a decreased activity in the aMCC and AI in the 'other' compared with the 'self' condition. This finding support the above-mentioned special sensitivity of schizophrenia patients to the affective states of others (Michaels and others 2014).

Generally, different fMRI studies using ToM tasks have shown an overall decrease in the neural activity of the core network (i.e., mPFC, PCC, and TPJ) in schizophrenia. For instance, Das and colleagues (2012) found that schizophrenia patients show decreased neural activity of the TPJ and inferior frontal gyrus (IFG) during a perspective-taking task using interacting geometric shapes, a modified paradigm described by Heider and Simmel (1944). Similarly, these patients showed decreased activity in the ventromedial prefrontal (vmPFC) and orbitofrontal (OFC) cortices when asked to identify objects based on the perspective of others (Eack and others 2013) as well as a decreased activity in the mPFC and TPJ during false-belief tasks (Dodell-Feder and others 2014; Lee and others 2011). However, an fMRI study by Brüne and colleagues (2008) found increased neural activity in the superior temporal gyrus (STG), dorsal mPFC and posteromedial cortices in a group of schizophrenia patients, although they performed similarly to controls in inferring others' intentions. This is in line with a recent study showing increased activity in the posterior STG and mPFC in schizophrenia patients compared with controls when inferring emotions from pictures of eyes (de Achával and others 2012), suggesting that greater neural activity is necessary for these patients to succeed in ToM tasks.

In sum, most behavioral and neuroimaging findings concur with a disturbance of schizophrenia patients in their abstract ability to infer others’ beliefs and intentions (i.e., a TT deficit), while some reports even described patients with enhanced capacities to automatically simulate the affective states of others (i.e., an ST excess).

Box 3: TT versus ST impairment in borderline personality disorder

Borderline personality disorder (BPD) patients have traditionally been linked to a dysfunction of behaviors that involve sharing the affective state of others (Fonagy 1991; Skodol 2007). However, Harari and others (2010) have recently shown in a behavioral experiment evaluating both ToM and empathy that BPD patients displayed alterations of abstract social-cognitive capacities similar to those of schizophrenia patients (see Box 2). That is, BPD patients were characterized by an impaired capacity to abstractly emulate the internal states of others (i.e., a TT deficit), as well as an increased tendency to empathize with the affective states of others (i.e., an ST excess). This is congruent with recent evidence showing that BPD patients can be as accurate (Preißler and others 2010; Schilling and others 2012) or even better than controls in perspective taking and other ToM tasks when emotional cues of the mental states of others are provided (Fertuck and others 2009; Frick and others 2012; Scott and others 2011). That is, though typically impaired to make inferences about others' mental states as captured by TT, these patients may compensate by means of a greater ability to automatically share their affective state, suggesting preserved ST capacities. Nevertheless, a recent study by Kalpakci and colleagues (2016) have challenged this view. These authors found the opposite pattern: adolescent BPD patients reported impaired empathy as well as enhanced ToM. Therefore, these contradictory behavioral findings cannot provide a clear mechanistic explanation of the BPD with respect to the TT versus ST frameworks of social cognition.

Neuroimaging findings in BPD patients have also provided largely conflicting results. In an fMRI study, Dziobek and colleagues (2011) found that these patients show decreased neural activity of the STG and increased activity in the middle insula compared with controls when asked to make inferences about others' affective states. These authors suggested that BPD patients might misinterpret others' emotional states in social interactions what, in turn, would lead to heightened distress and inappropriate own emotions (Dziobek and others 2011). However, Dinsdale and Crespi (2013) suggested that an enhanced capacity for the accurate identification of other people's emotional states in BPD patients has not been observed consistently and depends on the specific characteristics of the interaction situation.

In sum, both behavioral and neuroimaging findings offer mixed insights on the ability of the BPD patients to adequately understand the intentions of others by either making inferences about (i.e., TT) or reinstantiating (i.e., ST) their behavior.

Box 4: TT versus ST impairment in autism spectrum disorders

Early attempts to characterize social cognition deficits in the autism spectrum disorders (ASD) had focused on the abstract ability to infer the beliefs and intentions of others (Baron-Cohen and others 1985; Baron-Cohen and others 1995). More recently, it has been repeatedly highlighted that atypical empathic behavior in early childhood is a key feature of ASD (Baron-Cohen and Wheelwright 2004; Decety and Moriguchi 2007; Scambler and others 2007) and predicts later diagnosis (Ozonoff and others 2010). In contrast, different studies have showed that individuals with Asperger Syndrome preserve intact empathy capacities (Bird and others 2010; Dziobek and others 2008). Congruently, in a behavioral experiment using multi-dimensional empathy measures, Rogers and colleagues (2007) found no difference between Asperger Syndrome patients and healthy controls in empathy, while ToM abilities reported to be significantly impaired. Thus, the high variability across the spectrum of patients with autism have hindered previous attempts to characterize the social-cognitive impairments of these patients based on their ability to simulate and share the affective state of others (i.e., a ST deficit), their emulation capacity for the internal states of others (i.e., a TT deficit), or a combination of both processes (Baron-Cohen 2002; Baron-Cohen and others 2005; McIntosh and others 2006; Minio-Paluello and others 2009).

Many neuroimaging studies have reported that brain regions related to ToM (e.g., mPFC, TPJ, STS, and TP) consistently show decreased neural activity in ASD patients (Castelli and others 2002; Happe and others 1996; Wang and others 2006). Nevertheless, Wang and others (2007) found that explicit instructions to focus on the social stimuli can modulate the neural activity of the mPFC in these patients. That is, when the experimenters prompted ASD patients to attend to the facial expressions and tone of voice of the target, statistically significant differences could not be found anymore between the increased activity of the mPFC in the patient group compared with typically developed children. Yet, to the best of our knowledge, there is a scarcity of neuroimaging studies on affect sharing in ASD patients, preventing us from speculating about possible neural underpinnings of a social-cognitive impairment in these patients due to a ST dysfunction.

In sum, previous research at both the behavioral and neural levels corroborates an impaired capacity of ASD patients to make abstract inferences about the mental states of others that is in line with the TT framework. However, although there is additional behavioral evidence for a flawed ability to share the affective state of others, further supporting evidence from neuroimaging experiments is still needed.

References

Achim AM, Ouellet R, Roy M-A, Jackson PL. 2011. Assessment of empathy in first-episode psychosis and meta-analytic comparison with previous studies in schizophrenia. Psychiatry research 190(1):3-8.

Adolphs R. 1999. Social cognition and the human brain. Trends in cognitive sciences 3(12):469-479.

Adolphs R. 2003. Cognitive neuroscience of human social behaviour. Nature Reviews Neuroscience 4(3):165-178.

Apperly IA. 2008. Beyond Simulation–Theory and Theory–Theory: Why social cognitive neuroscience should use its own concepts to study “theory of mind”. Cognition 107(1):266-283.

Baillargeon R, Scott RM, Bian L. 2016. Psychological reasoning in infancy. Annual review of psychology 67:159-186.

Baker C, Saxe R, Tenenbaum J. Bayesian theory of mind: Modeling joint belief-desire attribution. Proceedings of the Cognitive Science Society; 2011.

Baron-Cohen S. 2002. The extreme male brain theory of autism. Trends in cognitive sciences 6(6):248-254.

Baron-Cohen S, Knickmeyer RC, Belmonte MK. 2005. Sex differences in the brain: implications for explaining autism. Science 310(5749):819-823.

Baron-Cohen S, Leslie AM, Frith U. 1985. Does the autistic child have a “theory of mind”? Cognition 21(1):37-46.

Baron-Cohen S, Wheelwright S. 2004. The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. Journal of autism and developmental disorders 34(2):163-175.

Baron-Cohen S, Campbell R, Karmiloff-Smith A, Grant J, Walker J. 1995. Are children with autism blind to the mentalistic significance of the eyes? British Journal of Developmental Psychology 13(4):379-398.

Baron-Cohen S, Wheelwright S, Hill J, Raste Y, Plumb I. 2001. The “Reading the Mind in the Eyes” test revised version: A study with normal adults, and adults with Asperger syndrome or high-functioning autism. Journal of child psychology and psychiatry 42(2):241-251.

Behrens TE, Hunt LT, Rushworth MF. 2009. The computation of social behavior. science 324(5931):1160-1164.

Bird G, Silani G, Brindley R, White S, Frith U, Singer T. 2010. Empathic brain responses in insula are modulated by levels of alexithymia but not autism. Brain:awq060.

Bloom P. 2016. Empathy and its discontents. Trends in cognitive sciences.

Bloom P. 2017. Empathy and Its Discontents. Trends in Cognitive Sciences 21(1):24-31.

Bora E, Yucel M, Pantelis C. 2009. Theory of mind impairment in schizophrenia: meta-analysis. Schizophrenia research 109(1):1-9.

Brass M, Schmitt RM, Spengler S, Gergely G. 2007. Investigating action understanding: inferential processes versus action simulation. Current Biology 17(24):2117-2121.

Brüne M, Lissek S, Fuchs N, Witthaus H, Peters S, Nicolas V, Juckel G, Tegenthoff M. 2008. An fMRI study of theory of mind in schizophrenic patients with “passivity” symptoms. Neuropsychologia 46(7):1992-2001.

Buccino G, Binkofski F, Fink GR, Fadiga L, Fogassi L, Gallese V, Seitz RJ, Zilles K, Rizzolatti G, Freund HJ. 2001. Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. European journal of neuroscience 13(2):400-404.

Buckner RL, Carroll DC. 2007. Self-projection and the brain. Trends in cognitive sciences 11(2):49-57.

Buttelmann D, Carpenter M, Tomasello M. 2009. Eighteen-month-old infants show false belief understanding in an active helping paradigm. Cognition 112(2):337-342.

Bzdok D, Schilbach L, Vogeley K, Schneider K, Laird AR, Langner R, Eickhoff SB. 2012. Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. Brain Structure and Function 217(4):783-796.

Call J, Tomasello M. 2008. Does the chimpanzee have a theory of mind? 30 years later. Trends in cognitive sciences 12(5):187-192.

Carruthers P. 1996. Simulation and self-knowledge: a defence of theory-theory. Theories of theories of mind:22-38.

Carruthers P. 2009. How we know our own minds: The relationship between mindreading and metacognition. Behavioral and brain sciences 32(02):121-138.

Carruthers P, Smith PK. 1996. Theories of theories of mind. Cambridge Univ Press.

Caspers S, Zilles K, Laird AR, Eickhoff SB. 2010. ALE meta-analysis of action observation and imitation in the human brain. Neuroimage 50(3):1148-1167.

Castelli F, Frith C, Happé F, Frith U. 2002. Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. Brain 125(8):1839-1849.

Cheng Y, Chen C, Lin C-P, Chou K-H, Decety J. 2010. Love hurts: an fMRI study. Neuroimage 51(2):923-929.

Chersi F, Ferrari PF, Fogassi L. 2011. Neuronal chains for actions in the parietal lobe: a computational model. PloS one 6(11):e27652.

Choi-Kain LW, Gunderson JG. 2008. Mentalization: ontogeny, assessment, and application in the treatment of borderline personality disorder. American Journal of Psychiatry 165(9):1127-1135.

Das P, Lagopoulos J, Coulston CM, Henderson AF, Malhi GS. 2012. Mentalizing impairment in schizophrenia: a functional MRI study. Schizophrenia research 134(2):158-164.

de Achával D, Villarreal MF, Costanzo EY, Douer J, Castro MN, Mora MC, Nemeroff CB, Chu E, Bär K-J, Guinjoan SM. 2012. Decreased activity in right-hemisphere structures involved in social cognition in siblings discordant for schizophrenia. Schizophrenia research 134(2):171-179.

De Lange FP, Spronk M, Willems RM, Toni I, Bekkering H. 2008. Complementary systems for understanding action intentions. Current biology 18(6):454-457.

De Vignemont F, Singer T. 2006. The empathic brain: how, when and why? Trends in cognitive sciences 10(10):435-441.

de Waal F. 1999. Good Natured: The Origins of Right and Wrong in Humans and Other Animals. Senior Managing Editor 6(1):63.

De Waal FB. 2008. Putting the altruism back into altruism: the evolution of empathy. Annu. Rev. Psychol. 59:279-300.

de Weerd H, Broers E, Verbrugge R. Savvy software agents can encourage the use of second-order theory of mind by negotiators. CogSci; 2015.

Decety J. 2010. To what extent is the experience of empathy mediated by shared neural circuits? Emotion Review 2(3):204-207.

Decety J, Chaminade T. 2003. Neural correlates of feeling sympathy. Neuropsychologia 41(2):127-138.

Decety J, Cowell JM. 2014. The complex relation between morality and empathy. Trends in cognitive sciences 18(7):337-339.

Decety J, Grèzes J. 1999. Neural mechanisms subserving the perception of human actions. Trends in cognitive sciences 3(5):172-178.

Decety J, Jackson PL. 2004. The functional architecture of human empathy. Behavioral and cognitive neuroscience reviews 3(2):71-100.

Decety J, Lamm C. 2006. Human empathy through the lens of social neuroscience. The Scientific World Journal 6:1146-1163.

Decety J, Moriguchi Y. 2007. The empathic brain and its dysfunction in psychiatric populations: Implications for intervention across different clinical conditions. BioPsychoSocial Medicine 1(1):1.

Devaine M, Hollard G, Daunizeau J. 2014. The social Bayesian brain: does mentalizing make a difference when we learn? PLoS computational biology 10(12):e1003992.

Di Pellegrino G, Fadiga L, Fogassi L, Gallese V, Rizzolatti G. 1992. Understanding motor events: a neurophysiological study. Experimental brain research 91(1):176-180.

Dimberg U, Öhman A. 1996. Behold the wrath: Psychophysiological responses to facial stimuli. Motivation and Emotion 20(2):149-182.

Dinsdale N, Crespi BJ. 2013. The borderline empathy paradox: evidence and conceptual models for empathic enhancements in borderline personality disorder. Journal of personality disorders 27(2):172.

Dodell-Feder D, Tully LM, Lincoln SH, Hooker CI. 2014. The neural basis of theory of mind and its relationship to social functioning and social anhedonia in individuals with schizophrenia. NeuroImage: Clinical 4:154-163.

Dziobek I, Preißler S, Grozdanovic Z, Heuser I, Heekeren HR, Roepke S. 2011. Neuronal correlates of altered empathy and social cognition in borderline personality disorder. Neuroimage 57(2):539-548.

Dziobek I, Rogers K, Fleck S, Bahnemann M, Heekeren HR, Wolf OT, Convit A. 2008. Dissociation of cognitive and emotional empathy in adults with Asperger syndrome using the Multifaceted Empathy Test (MET). Journal of autism and developmental disorders 38(3):464-473.

Eack SM, Wojtalik JA, Newhill CE, Keshavan MS, Phillips ML. 2013. Prefrontal cortical dysfunction during visual perspective-taking in schizophrenia. Schizophrenia research 150(2):491-497.

Eslinger PJ. 1998. Neurological and neuropsychological bases of empathy. European neurology 39(4):193-199.

Etzel JA, Gazzola V, Keysers C. 2008. Testing simulation theory with cross-modal multivariate classification of fMRI data. PloS one 3(11):e3690.

Fan Y, Duncan NW, de Greck M, Northoff G. 2011. Is there a core neural network in empathy? An fMRI based quantitative meta-analysis. Neuroscience & Biobehavioral Reviews 35(3):903-911.

Fertuck E, Jekal A, Song I, Wyman B, Morris M, Wilson S, Brodsky B, Stanley B. 2009. Enhanced 'Reading the Mind in the Eyes' in borderline personality disorder compared to healthy controls. Psychological medicine 39(12):1979-1988.

Fogassi L, Ferrari PF. 2007. Mirror neurons and the evolution of embodied language. Current directions in psychological science 16(3):136-141.

Fogassi L, Ferrari PF, Gesierich B, Rozzi S, Chersi F, Rizzolatti G. 2005. Parietal lobe: from action organization to intention understanding. Science 308(5722):662-667.

Fogassi L, Gallese V, Fadiga L, Rizzolatti G. Neurons responding to the sight of goal-directed hand/arm actions in the parietal area PF (7b) of the macaque monkey. Society of Neuroscience Abstracts; 1998.

Fonagy P. 1991. Thinking about thinking: Some clinical and theoretical considerations in the treatment of a borderline patient. The International journal of psycho-analysis 72(4):639.

Frick C, Lang S, Kotchoubey B, Sieswerda S, Dinu-Biringer R, Berger M, Veser S, Essig M, Barnow S. 2012. Hypersensitivity in borderline personality disorder during mindreading. PLoS One 7(8):e41650.

Frith CD, Frith U. 1999. Interacting minds--a biological basis. Science 286(5445):1692-1695.

Frith U, Frith CD. 2003. Development and neurophysiology of mentalizing. Philosophical Transactions of the Royal Society of London B: Biological Sciences 358(1431):459-473.

Frith U, Morton J, Leslie AM. 1991. The cognitive basis of a biological disorder: Autism. Trends in neurosciences 14(10):433-438.

Gallagher HL, Frith CD. 2003. Functional imaging of 'theory of mind'. Trends in cognitive sciences 7(2):77-83.

Gallagher S. 2001. The practice of mind. Theory, simulation or primary interaction? Journal of Consciousness Studies 8(5-6):83-108.

Gallagher S. 2008. Direct perception in the intersubjective context. Consciousness and Cognition 17(2):535-543.

Gallagher S, Hutto D. 2008. Understanding others through primary interaction and narrative practice. The shared mind: Perspectives on intersubjectivity:17-38.

Gallese V. 2003. The manifold nature of interpersonal relations: the quest for a common mechanism. Philosophical Transactions of the Royal Society of London B: Biological Sciences 358(1431):517-528.

Gallese V, Fadiga L, Fogassi L, Rizzolatti G. 1996. Action recognition in the premotor cortex. Brain 119(2):593-609.

Gallese V, Goldman A. 1998. Mirror neurons and the simulation theory of mind-reading. Trends in cognitive sciences 2(12):493-501.

Gallese V, Keysers C, Rizzolatti G. 2004. A unifying view of the basis of social cognition. Trends in cognitive sciences 8(9):396-403.

Gallese V, Sinigaglia C. 2011. What is so special about embodied simulation? Trends in cognitive sciences 15(11):512-519.

Gazzola V, Aziz-Zadeh L, Keysers C. 2006. Empathy and the somatotopic auditory mirror system in humans. Current biology 16(18):1824-1829.

Georgescu AL, Kuzmanovic B, Santos NS, Tepest R, Bente G, Tittgemeyer M, Vogeley K. 2014. Perceiving nonverbal behavior: Neural correlates of processing movement fluency and contingency in dyadic interactions. Human brain mapping 35(4):1362-1378.

Gillberg CL. 1992. The Emanuel Miller Memorial Lecture 1991. Autism and autistic-like conditions: subclasses among

disorders of empathy. Journal of Child Psychology and Psychiatry 33(5):813-842.

Goldman A. 2000. The mentalizing folk. In: Sperber D, editor. Metarepresentation. Vancouver Studies in Cognitive Science: Oxford University Press. p. 171-196.

Goldman AI. 1992. In defense of the simulation theory. Mind & Language 7(1-2):104-119.

Goldman AI. 2006. Simulating minds: The philosophy, psychology, and neuroscience of mindreading. Oxford University Press.

Gopnik A, Wellman HM. 1992. Why the child's theory of mind really is a theory. Mind & Language 7(1-2):145-171.

Gopnik A, Wellman HM. The theory theory. An earlier version of this chapter was presented at the Society for Research in Child Development Meeting, 1991.; 1994: Cambridge University Press.

Gordon RM. 1986. Folk psychology as simulation. Mind & Language 1(2):158-171.

Grèzes J, Frith C, Passingham RE. 2004. Inferring false beliefs from the actions of oneself and others: an fMRI study. Neuroimage 21(2):744-750.

Gusnard DA, Raichle ME. 2001. Searching for a baseline: functional imaging and the resting human brain. Nature Reviews Neuroscience 2(10):685-694.

Hampton AN, Bossaerts P, O'Doherty JP. 2008. Neural correlates of mentalizing-related computations during strategic interactions in humans. Proceedings of the National Academy of Sciences 105(18):6741-6746.

Happe F, Ehlers S, Fletcher P, Frith U, Johansson M, Gillberg C, Dolan R, Frackowiak R, Frith C. 1996. 'Theory of mind'in the brain. Evidence from a PET scan study of Asperger syndrome. Neuroreport 8(1):197-201.

Happé FG, Winner E, Brownell H. 1998. The getting of wisdom: Theory of mind in old age. Developmental psychology 34(2):358-362.

Harari H, Shamay-Tsoory SG, Ravid M, Levkovitz Y. 2010. Double dissociation between cognitive and affective empathy in borderline personality disorder. Psychiatry research 175(3):277-279.

Hare B, Call J, Agnetta B, Tomasello M. 2000. Chimpanzees know what conspecifics do and do not see. Animal Behaviour 59(4):771-785.

Hare B, Call J, Tomasello M. 2001. Do chimpanzees know what conspecifics know? Animal behaviour 61(1):139-151.

Hatfield E, Cacioppo JT, Rapson RL. 1993. Emotional contagion. Current directions in psychological science 2(3):96-100.

Hatfield E, Rapson RL, Le Y-CL. 2009. Emotional contagion and empathy.

Heider F, Simmel M. 1944. An experimental study of apparent behavior. The American Journal of Psychology 57(2):243-259.

Henry JD, Phillips LH, Ruffman T, Bailey PE. 2013. A meta-analytic review of age differences in theory of mind. American Psychological Association.

Horan WP, Jimenez AM, Lee J, Wynn JK, Eisenberger NI, Green MF. 2016. Pain empathy in schizophrenia: an fMRI study. Social cognitive and affective neuroscience:nsw002.

Hutto DD. 2008. The Narrative Practice Hypothesis: clarifications and implications. Philosophical Explorations 11(3):175-192.

Iacoboni M, Woods RP, Brass M, Bekkering H, Mazziotta JC, Rizzolatti G. 1999. Cortical mechanisms of human imitation. Science 286(5449):2526-2528.

Jacob P, Jeannerod M. 2005. The motor theory of social cognition: a critique. Trends in cognitive sciences 9(1):21-25.

Kalpakci A, Vanwoerden S, Elhai JD, Sharp C. 2016. The independent contributions of emotion dysregulation and hypermentalization to the “double dissociation” of affective and cognitive empathy in female adolescent inpatients with BPD. Journal of personality disorders 30(2):242-260.

Kanske P, Böckler A, Trautwein F-M, Lesemann FHP, Singer T. 2016. Are strong empathizers better mentalizers? Evidence for independence and interaction between the routes of social cognition. Social cognitive and affective neuroscience:nsw052.

Keysers C, Gazzola V. 2007. Integrating simulation and theory of mind: from self to social cognition. space 8:108-114.

Keysers C, Gazzola V. 2009. Expanding the mirror: vicarious activity for actions, emotions, and sensations. Current opinion in neurobiology 19(6):666-671.

Kilner JM. 2011. More than one pathway to action understanding. Trends in cognitive sciences 15(8):352-357.

Kohler E, Keysers C, Umilta MA, Fogassi L, Gallese V, Rizzolatti G. 2002. Hearing sounds, understanding actions: action representation in mirror neurons. Science 297(5582):846-848.

Koster-Hale J, Saxe R. 2013. Theory of mind: a neural prediction problem. Neuron 79(5):836-848.

Krupenye C, Kano F, Hirata S, Call J, Tomasello M. 2016. Great apes anticipate that other individuals will act according to false beliefs. Science 354(6308):110-114.

Lamm C, Decety J, Singer T. 2011. Meta-analytic evidence for common and distinct neural networks associated with directly experienced pain and empathy for pain. Neuroimage 54(3):2492-2502.

Lee J, Quintana J, Nori P, Green MF. 2011. Theory of mind in schizophrenia: exploring neural mechanisms of belief attribution. Social neuroscience 6(5-6):569-581.

Leslie AM. 1987. Pretense and representation: The origins of" theory of mind.". Psychological review 94(4):412.

Leslie AM. 1994. Pretending and believing: Issues in the theory of ToMM. Cognition 50(1-3):211-238.

Maylor EA, Moulson JM, Muncer AM, Taylor LA. 2002. Does performance on theory of mind tasks decline in old age? British Journal of Psychology 93(4):465-485.

McIntosh DN, Reichmann-Decker A, Winkielman P, Wilbarger JL. 2006. When the social mirror breaks: deficits in automatic, but not voluntary, mimicry of emotional facial expressions in autism. Developmental science 9(3):295-302.

Meltzoff AN, Moore MK. 1989. Imitation in newborn infants: Exploring the range of gestures imitated and the underlying mechanisms. Developmental psychology 25(6):954.

Michaels TM, Horan WP, Ginger EJ, Martinovich Z, Pinkham AE, Smith MJ. 2014. Cognitive empathy contributes to poor social functioning in schizophrenia: evidence from a new self-report measure of cognitive and affective empathy. Psychiatry research 220(3):803-810.

Minio-Paluello I, Baron-Cohen S, Avenanti A, Walsh V, Aglioti SM. 2009. Absence of embodied empathy during pain observation in Asperger syndrome. Biological psychiatry 65(1):55-62.

Mitchell JP. 2009. Social psychology as a natural kind. Trends in cognitive sciences 13(6):246-251.

Mitchell JP, Macrae CN, Banaji MR. 2006. Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. Neuron 50(4):655-663.

Mukamel R, Ekstrom AD, Kaplan J, Iacoboni M, Fried I. 2010. Single-neuron responses in humans during execution and observation of actions. Current biology 20(8):750-756.

Newen A. 2014. Understanding others: The person model theory. Open MIND. Open MIND. Frankfurt am Main: MIND Group.

Newen A, Schlicht T. 2009. Understanding other minds: A criticism of Goldman's Simulation Theory and an outline of the Person Model Theory. Grazer Philosophische Studien 79(1):209-242.

Nichols S, Stich SP. 2003. Mindreading: An integrated account of pretence, self-awareness, and understanding other minds. Clarendon Press/Oxford University Press.

Nishitani N, Hari R. 2000. Temporal dynamics of cortical representation for action. Proceedings of the National Academy of Sciences 97(2):913-918.

Ozonoff S, Iosif A-M, Baguio F, Cook IC, Hill MM, Hutman T, Rogers SJ, Rozga A, Sangha S, Sigman M. 2010. A prospective study of the emergence of early behavioral signs of autism. Journal of the American Academy of Child & Adolescent Psychiatry 49(3):256-266. e2.

Perner J. 1991. Understanding the representational mind. The MIT Press.

Perner J, Kühberger A. 2005. Mental Simulation. Other minds: How humans bridge the divide between self and others. The Guilfod Press New York. p. 174-189.

Pfeifer JH, Iacoboni M, Mazziotta JC, Dapretto M. 2008. Mirroring others' emotions relates to empathy and interpersonal competence in children. Neuroimage 39(4):2076-2085.

Piaget J. 1945. La formation du symbole chez l’enfant (Imitation, jeu et rêve, image et représentation) Col. Actualités pédagogiques et psychologiques, Delachaux et Niestlé Éd.

Preißler S, Dziobek I, Ritter K, Heekeren HR, Roepke S. 2010. Social cognition in borderline personality disorder: evidence for disturbed recognition of the emotions, thoughts, and intentions of others. Frontiers in behavioral neuroscience 4:182.

Premack D, Woodruff G. 1978. Does the chimpanzee have a theory of mind? Behavioral and brain sciences 1(04):515-526.

Preston SD, De Waal FB. 2002. Empathy: Its ultimate and proximate bases. Behavioral and brain sciences 25(1):1-20.

Raichle ME, MacLeod AM, Snyder AZ, Powers WJ, Gusnard DA, Shulman GL. 2001. A default mode of brain function. Proceedings of the National Academy of Sciences 98(2):676-682.

Ramnani N, Miall RC. 2004. A system in the human brain for predicting the actions of others. Nature neuroscience 7(1):85-90.

Rizzolatti G, Fadiga L, Gallese V, Fogassi L. 1996. Premotor cortex and the recognition of motor actions. Cognitive brain research 3(2):131-141.

Rizzolatti G, Fogassi L, Gallese V. 2001. Neurophysiological mechanisms underlying the understanding and imitation of action. Nature Reviews Neuroscience 2(9):661-670.

Roeyers H, Buysse A, Ponnet K, Pichal B. 2001. Advancing advanced mind-reading tests: empathic accuracy in adults with a pervasive developmental disorder. Journal of Child Psychology and Psychiatry 42(2):271-278.

Rogers K, Dziobek I, Hassenstab J, Wolf OT, Convit A. 2007. Who cares? Revisiting empathy in Asperger syndrome. Journal of autism and developmental disorders 37(4):709-715.

Ruby P, Decety J. 2004. How would you feel versus how do you think she would feel? A neuroimaging study of perspective-taking with social emotions. Journal of cognitive neuroscience 16(6):988-999.

Ruffman T. 1996. Do children understand the mind by means of simulation or a theory? Evidence from their understanding of inference. Mind & Language 11(4):388-414.

Santos NS, Kuzmanovic B, David N, Rotarska-Jagiela A, Eickhoff SB, Shah J, Fink GR, Bente G, Vogeley K. 2010. Animated brain: A functional neuroimaging study on animacy experience. NeuroImage 53(1):291-302.

Savla GN, Vella L, Armstrong CC, Penn DL, Twamley EW. 2012. Deficits in domains of social cognition in schizophrenia: a meta-analysis of the empirical evidence. Schizophrenia bulletin:sbs080.

Saxe R. 2005. Against simulation: the argument from error. Trends in cognitive sciences 9(4):174-179.

Saxe R, Carey S, Kanwisher N. 2004. Understanding other minds: linking developmental psychology and functional neuroimaging. Annu. Rev. Psychol. 55:87-124.

Saxe R, Kanwisher N. 2003. People thinking about thinking people: the role of the temporo-parietal junction in “theory of mind”. Neuroimage 19(4):1835-1842.

Scambler D, Hepburn S, Rutherford M, Wehner E, Rogers SJ. 2007. Emotional responsivity in children with autism, children with other developmental disabilities, and children with typical development. Journal of autism and developmental disorders 37(3):553-563.

Schilbach L, Eickhoff SB, Rotarska-Jagiela A, Fink GR, Vogeley K. 2008. Minds at rest? Social cognition as the default mode of cognizing and its putative relationship to the “default system” of the brain. Consciousness and cognition 17(2):457-467.

Schilling L, Wingenfeld K, Löwe B, Moritz S, Terfehr K, Köther U, Spitzer C. 2012. Normal mind-reading capacity but higher response confidence in borderline personality disorder patients. Psychiatry and Clinical Neurosciences 66(4):322-327.

Scott LN, Levy KN, Adams Jr RB, Stevenson MT. 2011. Mental state decoding abilities in young adults with borderline personality disorder traits. Personality Disorders: Theory, Research, and Treatment 2(2):98.

Scott RM, Baillargeon R. 2017. Early False-Belief Understanding. Trends in Cognitive Sciences.

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

Shamay-Tsoory SG, Aharon-Peretz J, Perry D. 2009. Two systems for empathy: a double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial prefrontal lesions. Brain 132(3):617-627.

Shamay-Tsoory SG, Tomer R, Berger B, Aharon-Peretz J. 2003. Characterization of empathy deficits following prefrontal brain damage: the role of the right ventromedial prefrontal cortex. Journal of cognitive neuroscience 15(3):324-337.

Singer T, Lamm C. 2009. The social neuroscience of empathy. Annals of the New York Academy of Sciences 1156(1):81-96.

Singer T, Leiberg S. 2009. Sharing the emotions of others: The neural bases of empathy. The cognitive neurosciences IV. MIT Press. p. 971–984.

Singer T, Seymour B, O'Doherty J, Kaube H, Dolan RJ, Frith CD. 2004. Empathy for pain involves the affective but not sensory components of pain. Science 303(5661):1157-1162.

Skodol AE. 2007. Borderline personality as a self-other representational disturbance. Journal of personality disorders 21(5):500.

Spreng RN, Mar RA, Kim AS. 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.

Sprong M, Schothorst P, Vos E, Hox J, Van Engeland H. 2007. Theory of mind in schizophrenia. The British Journal of Psychiatry 191(1):5-13.

Tomasello M, Call J, Hare B. 2003. Chimpanzees versus humans: it´ s not that simple. Trends in Cognitive Sciences 7(6):239-240.

Uddin LQ, Iacoboni M, Lange C, Keenan JP. 2007. The self and social cognition: the role of cortical midline structures and mirror neurons. Trends in cognitive sciences 11(4):153-157.

Umilta MA, Kohler E, Gallese V, Fogassi L, Fadiga L, Keysers C, Rizzolatti G. 2001. I know what you are doing: A neurophysiological study. Neuron 31(1):155-165.

Van Overwalle F. 2009. Social cognition and the brain: a meta-analysis. Human brain mapping 30(3):829-858.

Van Overwalle F, Baetens K. 2009. Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. Neuroimage 48(3):564-584.

Vogeley K. 2017. Two social brains: neural mechanisms of intersubjectivity. Phil. Trans. R. Soc. B 372(1727):20160245.

Vogeley K, Bussfeld P, Newen A, Herrmann S, Happé F, Falkai P, Maier W, Shah NJ, Fink GR, Zilles K. 2001. Mind reading: neural mechanisms of theory of mind and self-perspective. Neuroimage 14(1):170-181.

Vogeley K, Fink GR. 2003. Neural correlates of the first-person-perspective. Trends in cognitive sciences 7(1):38-42.

Vogeley K, May M, Ritzl A, Falkai P, Zilles K, Fink GR. 2004. Neural correlates of first-person perspective as one constituent of human self-consciousness. Journal of cognitive neuroscience 16(5):817-827.

Vogeley K, Roepstorff A. 2009. Contextualising culture and social cognition. Trends in cognitive sciences 13(12):511-516.

Wang AT, Lee SS, Sigman M, Dapretto M. 2006. Neural basis of irony comprehension in children with autism: the role of prosody and context. Brain 129(4):932-943.

Wang AT, Lee SS, Sigman M, Dapretto M. 2007. Reading affect in the face and voice: neural correlates of interpreting communicative intent in children and adolescents with autism spectrum disorders. Archives of general psychiatry 64(6):698-708.

Wellman HM. 1990. The child's theory of mind.

Wellman HM, Cross D, Watson J. 2001. Meta-analysis of theory-of-mind development: the truth about false belief. Child development 72(3):655-684.

Wimmer H, 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.

Yoshida W, Dolan RJ, Friston KJ. 2008. Game theory of mind. PLoS computational biology 4(12):e1000254.

Zahn-Waxler C, Robinson JL, Emde RN. 1992. The development of empathy in twins. Developmental psychology 28(6):1038.

Zaki J, Ochsner KN. 2012. The neuroscience of empathy: progress, pitfalls and promise. Nature neuroscience 15(5):675-680.