Reframing Social Cognition:

Relational versus Representational Mentalising

Accepted July 7th, 2020, at Psychological Bulletin.

Eliane Deschrijver

Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, 9000, Ghent, Belgium

School of Psychology, University of New South Wales (UNSW) Sydney, New South Wales 2052, Australia

e.deschrijver@unsw.edu.au

Colin Palmer

School of Psychology, UNSW Sydney, New South Wales 2052, Australia

colin.palmer@unsw.edu.au

Funding. E. D. received funding for the *AutiBelieve* project from the Research Foundation Flanders (FWO, postdoctoral fellowship).

Author contributions. E.D. initiated and developed the *AutiBelieve* project, and secured project-related funding. E. D. and C.P. were involved in theoretical discussions. E. D. drafted, finalized and submitted the manuscript. Both authors contributed to the writing and revisions of the paper and approved the final version of the manuscript for submission.

Acknowledgements. The authors are thankful to Prisca Bauer, Andrew Sims, Emiel Cracco, and Laura Mathys for comments on earlier versions of the manuscript, to Colin Clifford for providing a welcoming work environment for this project.

Conflicts of interest. The authors declare that they have no competing interests.

Abstract

The most dominant theory of human social cognition, the Theory of Mind hypothesis, emphasises our ability to infer the mental states of others. After having represented the mental states of another person, however, we can also have an idea of how well our thinking aligns with theirs, and our sensitivity to this alignment may guide the flow of our social interactions. Here, we focus on the distinction between 'mindreading' (inferring another's mental representation) and detecting the extent to which a represented mental state of another person is matching or mismatching with our own (mental conflict monitoring). We propose a reframing for mentalising data of the past 40 years in terms of mental conflict monitoring rather than mental representation. Via a systematic review of 51 false belief neuroimaging studies, we argue that key brain regions implicated in false belief designs (namely, temporoparietal junction areas) may methodologically be tied to mental conflict rather than to mental representation. Patterns of false belief data suggests that autism may be tied to a subtle issue with monitoring mental conflict combined with intact mental representation, rather than to lacking mental representation abilities or 'mindblindness' altogether. The consequences of this view for the larger social-cognitive domain are explored, including for perspective taking, moral judgements, and understanding irony and humour. This provides a potential shift in perspective for psychological science, its neuroscientific bases, and related disciplines: Throughout life, an adequate sensitivity to how others think differently (relational mentalising) may be more fundamental to navigating the social world than inferring which thoughts others have (representational mentalising).

Keywords: Theory of Mind, temporoparietal junction, autism, false belief, self-other distinction.

Public significant statement: This review synthesises the data on human social cognition and argues for a central role of mental conflict monitoring rather than 'mindreading' in social development and interaction. It suggests that people on the autism spectrum may be well able to grasp *what* others think, in contrast to popular and scientific belief, while they may experience more subtle issues after this with monitoring the extent to which others are *thinking differently from themselves*. Throughout life, an adequate processing of the extent to which others are on the same page may be more fundamental to navigating the social world than inferring mental states of others. A lay summary can also be found in the Section 'Relational mentalising in everyday life'.

After four decades of research into Theory of Mind, neuroscientists and psychologists have come to accept that inferring the content of others' mental states or 'mindreading' may be the human brain's foremost activity when interacting with other people (Baron-Cohen, Leslie, & Frith, 1985). It has been similarly argued that key regions in the brain's 'social' network, namely the temporoparietal junction (TPJ) areas, are specifically dedicated to representing others' mental states (Apperly, Samson, Chiavarino, & Humphreys, 2004; Samson, Apperly, Chiavarino, & Humphreys, 2004; Saxe & Kanwisher, 2003; Saxe & Powell, 2006; Saxe & Wexler, 2005a). "Mindblindness", a deficiency in mental representation, has been one of the primary explanations behind social differences, particularly in autism (Alcalá-López, Vogeley, Binkofski, & Bzdok, 2019; Baron-Cohen et al., 1985; Lombardo, Chakrabarti, Bullmore, & Baron-Cohen, 2011; Senju, Southgate, White, & Frith, 2009). In social interactions, however, it is also crucial to be sensitive to the extent that others *share* one's own understanding of the world, even if one grasps the content of the other's thinking already. In any conversation, the other person's mental perspective of the world will most likely not fully align with ours. This lack of alignment may be important to guiding the flow of our interaction with them. Instead of '*computing what others think*', does focussing on the notion of '*understanding when others think differently'* help to unravel the mechanisms of social cognition?

In this theoretical review, we focus on the distinction between *inferring the content of another's mental state* (henceforth: inferring an other-related mental representation) and monitoring the extent to which another's mental state representation *is mismatching with one's own* (henceforth: monitoring mental conflict). Importantly, an analogous distinction between representation and conflict monitoring on the basis of own- and other-related representations has been fundamental to driving innovation in experimental designs and conceptual progress

in the domain of action perception, the other main domain of human social-cognitive research. In this field, scientific consensus exists that the *representation* of others' actions and *monitoring conflict* between self- and other-related actions are cognitive mechanisms investigated by means of critically different experimental designs. In particular, comparing a socially *incongruent* versus *congruent* condition is used to isolate social conflict monitoring processes. In spite of strong empirical, methodological and theoretical links between the action perception and Theory of Mind domains, the latter has not yet developed a similar theoretical and empirical tradition of understanding socially incongruent (versus congruent) data on the level of social conflict monitoring rather than social representation. We will point out that this interpretational preference bears on a seminal logical argument about how to establish evidence for mental representation, which doesn't account for how social conflict occurring after mental states are represented in the brain may influence experimental outcomes. We argue that behavioural and neuroimaging data derived from experimental designs commonly used to investigate Theory of Mind are better interpreted in terms of mental conflict monitoring (in a relational framework) rather than in terms of inferring mental representations (in a representational framework), particularly when the appropriate experimental contrasts are used. In this light, we scrutinise the neuroimaging evidence that bears on how 'the contents of other people's mental states' are represented in the brain (namely, in temporoparietal junction areas; TPJ): The experimental design used in many imaging studies, we argue, may better isolate the neural signature of mental conflict monitoring instead.

With autism as a clinical test case, we then review the evidence of how mental conflict monitoring and mental representation are tied to real-world social experience. The Theory of Mind hypothesis of autism (Baron-Cohen et al., 1985), which asserts that individuals on the

autism spectrum have lacking mental representations, has been the most dominant account for understanding social difficulties in the past thirty-five years. Centred on the methodological arguments that we outline, we argue that the pattern of results across studies using different dependent measures better fit an interpretation where individuals on the autism spectrum have difficulty coping more specifically when their representation of another person's mental state *diverges* from their own, with other-related mental representation as such being intact. We review evidence that bears on the question of whether mentalising abilities are present in the neurotypical brain from early in life, while being affected in the autistic brain around the same developmental period, and make a similar point about the study of mentalising processes in non-human primates.

In the final section of the paper, we explore the implications of this framework for the methodological designs used in the larger social-cognitive domain, contrasting the fields of action perception, empathy, and the observation of touch (which initially developed representation-only paradigms) against the fields of mentalising, perspective taking, irony, humour, sarcasm, lie detection and moral dilemmas (which relied initially on social conflict paradigms). Understanding results under a relational rather than a representational framework helps to bring clarity to the data in these fields, and may be essential to defining mechanisms of social cognition that are shared across domains. Overall, we argue that data patterns over a range of methods (neuroimaging, eyetracking, verbal responses, reaction times (RTs)), populations (neurotypical, clinical, animal and developmental), and areas of social cognition (Theory of Mind, perspective taking, moral decision making, and others) may be interpreted more parsimoniously in light of an appropriate distinction between relational and representational social cognition.

In sum, by emphasising an interactive mechanism of *relating* others' mental state to ours (*relational* mentalising) rather than a mechanism of attributing *content* to them (*representational* mentalising), we aim to introduce a shift in perspective for psychological science and related disciplines: The essential ingredient for neurotypical social cognition may lie in monitoring alignment with others rather than in inferring social representations, mindreading or Theory of Mind as such.

Reframing the Interpretation of Data in the Theory of Mind Domain

Theory of Mind was initially introduced (Premack & Woodruff, 1978) as the ability to impute mental states to oneself and others. This definition, which is to date still the most widespread, gives weight to what we refer to as inferring a mental representation. The most popular experimental design for assessing Theory of Mind abilities is called the false belief task. For instance in the Sally-Anne variant of this task (Baron-Cohen et al., 1985), participants observe Sally placing an object (typically, a ball) in a box, before leaving the scene. After this, Anne moves the ball to a basket and Sally returns. The participant is then asked where Sally will look for the ball. Participants with well-developed Theory of Mind abilities succeed in predicting Sally's ball-searching behaviour based on her (false) belief about the ball's location. Henceforth, we will refer to all tasks centred on the Sally-Anne design (e.g., where an agent hides an object, which is in absence relocated) as well as designs that use false belief manipulations more loosely based on those used in the Sally Anne task (e.g., Saxe & Kanwisher, 2003; Saxe & Wexler, 2005) as false belief tasks. We describe mentalising tasks using still other designs (e.g. Castelli, Happé, Frith, & Frith, 2000; Dziobek et al., 2006) in the Section 'Relational Mentalising: Mirroring Others' Mental States?'.

Historically, the presence of mental conflict in false belief tasks (i.e., the fact that the beliefs of Sally and of the participant about where the ball is located are mismatching) was considered a methodological necessity or even a methodological artefact for the primary goal of understanding individuals' abilities for representation of others' mental states. Without it, it was reasoned, a participant's verbal responding would not be informative about their ability to *represent* other people's mental states (Dennett, 1978): An expression of Sally's behaviour based on her belief wouldn't be distinguishable from an expression of the participant's own belief or perception of the world. Consequently, empirical data in mentalising tasks are typically interpreted in terms of the participants' abilities to *represent* the content of another's mental state, i.e., the participant does or does not understand what the other person thinks (e.g., Alcalá-López, Vogeley, Binkofski, & Bzdok, 2019; Baron-Cohen et al., 1985; Lombardo et al., 2007; Lombardo, Chakrabarti, Bullmore, & Baron-Cohen, 2011; Overwalle, 2009; Premack & Woodruff, 1978; Saxe & Kanwisher, 2003; Saxe & Powell, 2006). This has contributed more broadly to mindreading or inferring a mental representation being seen as the core of human social cognition.

We will illustrate that interpreting false belief data in terms of mental conflict monitoring instead of in terms of representation (or in terms of both) may prove crucial to understanding data patterns in the mentalising domain, and the mechanisms of social cognition more broadly. This puts the weight on mental conflict monitoring rather than mental representation as the core cognitive process under examination.

This reframing of data in the mentalising domain leads to three key advances:

First, the most dominant interpretation of the role of the TPJ in mentalising research is

mental representation. In contrast, the methodological argument we present suggests that the TPJ may be tied specifically to mental conflict monitoring in the false belief tasks commonly used to study Theory of Mind. In addition, we put forward the possibility that TPJ activation in other mentalising paradigms may be tied specifically to mental conflict monitoring as well.

Second, the common interpretation of false belief data in terms of mental representation has led to the widespread assumption that an atypical effect in false belief conditions necessarily signifies lacking mental *representations*. We explore how variation in the patterns of performance across different dependent measures of false belief tasks may signify a disturbed monitoring of mental *misalignment* after having represented mental states, not an insensitivity to a mental state per se. This is particularly important for interpreting data focusing on populations on the autism spectrum, where data in false belief conditions are commonly interpreted as evidence for a lack of mental representations or *mindblindness* (e.g., Senju, Southgate, White, & Frith, 2009). We outline a similar argument with respect to lacking effects in developmental and non-human primate data.

Third, social conflict tasks are commonly interpreted in terms of representational processes in some scientific fields (e.g., the Theory of Mind domain, lie detection, humour understanding, moral dilemmas, irony, sarcasm and perspective taking domain) and relational processes in other fields (e.g. the action perception domain, touch and empathy domain). This conceptual difference may be the consequence mostly of a methodological limitation in the former domains, not of actual phenomenological differences necessarily. We discuss how aligning these fields in how they interpret comparable social conflict tasks may help to advance our understanding of shared mechanisms across different domains of social cognition.

Representational Versus Relational Interpretations in the Action Perception Domain: A Critical Distinction in Designs

With the discovery of the mirror neuron system (Rizzolatti & Craighero, 2004; Rizzolatti, Fogassi, & Gallese, 2001), researchers started to focus on where and how we may neurally *represent* actions that we observe others performing. Human fMRI studies revealed that brain areas implicated in representing one's own actions such as the premotor cortex and somatosensory cortices are also active when observing other people's actions (Gazzola & Keysers, 2009; Keysers et al., 2004; Keysers, Kaas, & Gazzola, 2010). Our brain is thought to 'mirror' motor as well as tactile aspects of what others experience while they are acting (Iacoboni & Dapretto, 2006; Rizzolatti & Craighero, 2004). This so-called neural imitation of actions was initially emphasised to be a central mechanism for facilitating an intuitive understanding of others, such as in understanding their goals and intentions (Rizzolatti & Craighero, 2004). Correspondingly, it was hypothesised that *deficits* in the representation of others' actions could lead to social difficulties in clinical conditions such as autism (known as the 'broken mirror neuron' theory of autism; e.g., Williams, Whiten, & Singh, 2004; see Figure 1). Thereafter, it was found that the human brain experiences *action interference* when trying to perform an action (e.g., a lifting motion of the index finger) that mismatches with an action that we simultaneously observe another person performing (e.g., a lifting motion of the middle finger; Brass, Bekkering, & Prinz, 2001; Brass, Bekkering, Wohlschläger, & Prinz, 2000; Brass, Zysset, & von Cramon, 2001; for a recent meta-analysis on automatic imitation, see Cracco et al., 2018). Importantly, it was shown that the human brain engages the TPJ when observing an action that mismatches one's own, establishing a neural mechanism for action conflict monitoring (Brass, Derrfuss, & Von Cramon, 2005; Spengler, Bird, & Brass, 2010).

Action *representation* studies compare a condition where another person's action is visible with a baseline condition where no (human) movement is visible, with the participant themself performing no action in either case. Such a baseline condition can consist, for instance, of viewing similar mechanical movements (e.g., bouncing balls; Oberman, Ramachandran, & Pineda, 2008), static images of body parts (Gazzola & Keysers, 2009; Martineau, Andersson, Barthélémy, Cottier, & Destrieux, 2010), or scrambled images of body parts (Gazzola & Keysers, 2009). This design deliberately reveals neural activity or behavioural differences specifically tied to *representing* the observed action. An atypical effect in the action observation versus the control condition, instead, would indicate an issue with action representations in the brain of the observer. It is crucial to understand, in contrast, that studies of action *conflict* have methodologically always contrasted a socially incongruent condition (i.e., where there is misalignment between our own action and the action we observe another performing) with a socially congruent condition (i.e., where our own action is identical to the one we observe). In a balanced design, the specific nature of the action (e.g., lifting index finger vs. lifting middle finger) is also controlled across the contrast (Brass et al., 2000). For example, a balanced design might compare two 'incongruent action' conditions (I perform action A while observing action B; or, I perform action B while observing action A) to two 'congruent action' conditions (we both perform action A; or, we both perform action B). The comparison leaves behind processes related specifically to the *conflict* between one's own and others' actions: Such effects are strictly speaking not directly informative about whether the individual *represents* the other-related action, as the difference between both conditions consist of a change in alignment between own and other-related actions, not a particular action as such. For example, larger RTs and error rates in the socially incongruent versus the congruent condition are described in terms of the individual's ability to deal with the conflict inherent in the former condition (Deschrijver, Wiersema, & Brass, 2017;

Deschrijver, Wiersema, & Brass, 2017), and the TPJ activity found in this contrast is tied to an action conflict monitoring mechanism (Brass et al., 2005; Spengler, Bird, et al., 2010). Likewise, an atypical incongruent versus congruent effect in such a paradigm (e.g., larger RTs or less TPJ activity) signifies a lesser monitoring of the misalignment between the two conditions, not an absence of action representation per se. Thus, separate experimental paradigms are used to isolate action representation and action conflict monitoring (see Figure 1, left), ensuring that the methodological focus of the designs align with the interpretational one. With this approach, the action perception domain has achieved differentiated knowledge on how the human brain respectively processes action conflict (i.e., in the TPJ) versus how it represents others' actions per se (i.e., in the mirror neuron system).

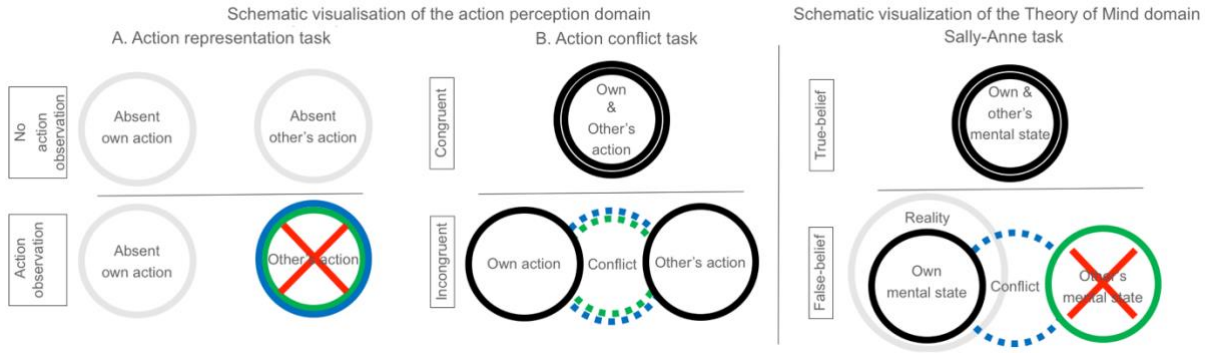

Figure 1. Left. Schematic representation of the action perception domain. Action representations are symbolised by full circles, in which a light grey circle signifies an absence of an action representation. Dotted circles signify the presence of conflict. The most common interpretational focus of the respective tasks is indicated in green, their methodological focus in blue. A. In the action representation paradigm, no conflict is elicited between the representations (circles) in the experimental action observation condition, as only an other-related but no own action representation should be present in the (pre)motor cortex of the observer. The control condition consists of a baseline condition where no action is observed nor executed by the participant: There shouldn't be any own- or other-related action representations present in the (pre)motor cortex of the participant. The results of the contrast yields activation related to the other-related action. Within this domain, the methodological focus is carefully fit to the most common interpretational focus, as both focus on the neural representation of other's actions. The 'broken' mirror hypothesis of autism, which proposes an absence of other-related neural representations in autism, is symbolised in red. *B.* In the action conflict task, conflict is implemented between an

own- and an other-related action in the incongruent experimental condition. The control condition, the congruent action condition, consists of aligning own- and other-related representations, yielding no action conflict. The result of the contrast yields activation related to action conflict only. The methodological focus of this task thus aligns with its interpretational one, as both focus on action conflict, not on action representation. *Right*. Schematic representation of the false belief domain. The design of the task is similar to the action conflict task: Diverging own and other-related mental representations are compared to aligning ones, if a true belief condition is used as a control. Under a traditional representational interpretation framework (see Table 1), the methodological and interpretational foci in Theory of Mind are not aligned: The experimental manipulation implements social conflict, whereas the data are mainly interpreted in terms of other-related representation abilities. The Theory of Mind hypothesis of autism, which asserts lacking other-related representation, is symbolised in red: Scholars have most commonly interpreted atypical effects in this task in terms of lacking other-related mental representation.

A Representational Versus Relational Interpretation of Social Conflict Data in the False Belief Design

From its inception, the understanding of Theory of Mind has leaned towards an interpretation primarily in terms of mental representation (Baron-Cohen et al., 1985; Premack & Woodruff, 1978). In fact, it is common for data in false belief tasks to be interpreted without reference to the role of social conflict. This includes nonverbal measures in false belief tasks, such as neuroimaging, behavioural or looking time measures. For instance, brain activity in TPJ areas in false belief conditions is typically interpreted as signifying the extent to which a participant *represents* another person's mental state. Reaction time and looking time differences in conditions where the other's mental state misaligns with the participant's mental state are similarly explained as signifying the extent to which the person represents the other mental state (i.e., other-related mental representation), not as showing the extent to which a person is processing social conflict (e.g., Gliga, Senju, Charman, & Johnson, 2014; Kovács,

Téglás, & Endress, 2010; Senju et al., 2009). Methodologically, however, the representation of others (i.e., Sally and I *have* a particular mental state) and social conflict monitoring (i.e., Sally has a mental state that is *conflicting* with mine) have been significantly more entangled in a false belief paradigm than in an action representation task. Like action conflict tasks, false belief tasks can be thought of as *social congruency* designs: The conflict between one's own mental states and another person's mental states is analogous to the conflict between one's own actions and another person's actions. Hence, the methodological focus of false belief tasks, which is mental conflict, currently does not correspond with the typical interpretational one, which is mental representation (see Figure 1, right).

As described in the previous Section, congruency designs allow foremost for gaining knowledge of *social conflict processes* when a condition is implemented to control for representation per se. Methodologically speaking, such a control condition should consist of *aligning* instead of *conflicting* self- and other-related representations, so that the difference between the two entails a *change in alignment*, not a particular mental state per se. Notably, some studies of false belief tasks have implemented such a control condition, which can be termed the 'true belief condition' in contrast to the primary 'false belief condition'. The true belief condition involves observing another person (e.g., Sally) who has a belief about an object's location that is identical to the participant's own belief. This experimental design bears an obvious similarity with the incongruent versus congruent contrast used in action conflict tasks (Brass et al., 2000; Brass, Zysset, et al., 2001). Specifically, the resulting 'false belief' versus 'true belief' contrast compares a socially mismatching condition (e.g., where the other person's mental representation of the situation misaligns with one's own) with a socially matching condition (i.e., where both align). If own- and other-related mental state representations are present in both conditions (see further in this Section), the comparison

should be expected to yield insights into the extent to which the brain is processing the conflict that is present in the false belief condition, as this is then the only aspect that differs between both conditions. This is particularly the case when a balanced design is implemented, which controls the specific content of the beliefs across the contrast. For example, a balanced design might compare two 'false belief' conditions (I believe A and Sally believes B; or, I believe B and Sally believes A) to two 'true belief' conditions (Sally and I both believe A; or, we both believe B; e.g., Bardi, Desmet, Nijhof, Wiersema, & Brass, 2016; Deschrijver, Bardi, Wiersema, & Brass, 2016; Kovács, Téglás, & Endress, 2010; Nijhof, Bardi, Brass, & Wiersema, 2018).

Why then is data in false belief tasks commonly interpreted in terms of mental representation rather than mental conflict, despite the use of a social congruency design? One important limitation that mentalising researchers face when trying to assess whether an individual possesses the ability to represent other's mental states, is that the participant will have their own mental states which may be used to solve the task. So when Premack and Woodruff (1978) proposed Theory of Mind as the ability to impute mental states to oneself and others, they aimed to assess whether chimpanzees have an expectation of another's behaviour (e.g., when observing an individual trying to solve a problem) that is *better* explained by mental state representation than from their own habits of thought or knowledge about the world (e.g., experienced behavioural regularities in how others act). In a seminal argument, Dennett (1978) argued that designs in which the other's mental state corresponds to reality (i.e., in a true belief situation) cannot entirely rule out such alternative explanations: In such situations, it was reasoned, you may attribute to them awareness of your model of the world without necessarily generating a representation of mental states per se, either yours or theirs (e.g., Dennett, 1978; Martin & Santos, 2016; Phillips & Norby, 2019). When the mental

state of another person *diverges* from their own model of reality (i.e., a false belief situation), instead, you need to do the work of imagining what model of the world would have been built by the other's partial perception of events (i.e., Sally didn't perceive the object moving to another location), distinct from one's own model of the world. Other-related mental representation was seen as prompted by, but not equivalent to, detecting the conflict between one's own and the other's perceptual history or awareness (henceforth: perceptual conflict).

It appeared that a number of populations show atypical results in false belief tasks even if clearly able to understand true belief situations, for instance young children (Priewasser, Rafetseder, Gargitter, & Perner, 2018; Ruffman, 1996; Sodian & Thoermer, 2008; Surian, Caldi, & Sperber, 2007), and non-human primates (Hare, Call, & Tomasello, 2001; Marticorena, Ruiz, Mukerji, Goddu, & Santos, 2011; for a summary, see Martin & Santos, 2016). This confirmed for mentalising scholars an intuition that generating someone else's specific representation of the world (i.e., an other-related mental representation) while understanding a false belief is a more difficult and qualitatively different process from attributing to them the one you have already have, for instance, in a true belief situation (Hare et al., 2001; Marticorena et al., 2011; Martin & Santos, 2014, 2016; Rothmayr et al., 2011a; Sodian & Thoermer, 2008; Surian et al., 2007). The Theory of Mind domain thus settled on a dominant conceptual interpretation of the socially incongruent (i.e., the false belief) condition mainly in terms of other-related representation (see Figure 1, right). From this point of view, a false versus true belief comparison is in essence not very different from the methodological comparison used in an action observation paradigm: In all dependent measures, the experimental condition would yield insights in other-related representation abilities (prompted by perceptual conflict), with the control condition accounting for non-representational processes. Because of this, the field subsequently went great lengths to develop control

conditions for the false belief condition that may account for perceptual conflict monitoring. Henceforth, we will refer to the lines of thought we just discussed as the 'representational mentalising framework (see Table 1).

Within the action perception domain, the idea that a socially congruent condition (like the true belief condition) may not evoke actual other-related representation processes has never been on the table (Brass, Bekkering, et al., 2001; Brass et al., 2000, 2005; Brass, Zysset, et al., 2001; Spengler, Von Cramon, & Brass, 2009). This seems in part a consequence of this domain not being subjected to the aforementioned methodological limitation: A participant's own action representations can be easily controlled when investigating how we represent another's action (i.e., the experimenter makes sure the participant does not move while observing this action). The domain thus evidenced early in its history that the human brain presumably always mirrors observed actions, regardless of any social conflict (Gazzola & Keysers, 2009; Iacoboni & Dapretto, 2006; Keysers et al., 2004; Rizzolatti & Craighero, 2004; Rizzolatti et al., 2001). In fact, conflict monitoring *after* representational processes have taken place is thought to exist precisely because we seem to mirror others' actions all the time: We represent others' actions even when they may hinder own action execution. As a consequence, scholars prefer to use the congruent action condition to isolate conflict monitoring from in nearly all of its studies (for a recent meta-analysis and review see Cracco et al., 2018; Heyes, 2011). In sum, the action perception domain adopted what we refer to as a relational framework for incongruent action condition interpretation (see Table 1).

Distinct conceptual interpretations in different social-cognitive domains for what is essentially the same social congruency design is in principle not a problem if the characteristics of the psychological processes involved are known to be fundamentally

different. However, there are reasons to think that the processes involved in mentalising and action conflict tasks are not. Because performance in action conflict tasks has been found to be correlated with TPJ activity while participants engage in mental state attribution (Brass et al., 2009; Spengler et al., 2010), it has been argued that a shared mechanism may serve them both. Neuromodulation (Hogeveen et al., 2014; Nobusako, Nishi, Nishi, Shuto, & Asano, 2017; Santiesteban, Banissy, Catmur, & Bird, 2015; Santiesteban, Banissy, et al., 2012; Sowden, Wright, et al., 2015), lesion data (Spengler, von Cramon, & Brass, 2010), clinical data (Spengler, Bird, et al., 2010) and training paradigms (Santiesteban, White, et al., 2012) amongst others have yielded additional evidence for a link between the action conflict and the mentalising domain (as well as other domains), often identified as located in the TPJ area. For this reason, the TPJ was hypothesised to host a social conflict monitoring mechanism common to these domains (Brass et al., 2009; Spengler, Bird, et al., 2010; Spengler et al., 2009) which, if it fulfils the same role across domains, should detect and solve the conflict that may arise after one represents an other-related representation next to an own-related representation irrespective of whether they align.

If applied to the mentalising domain, this is a premise diametrically opposing the idea that social conflict detection *precedes* other-related representation, and that socially aligning others (e.g., in a true belief situation) may not be represented at all (Dennett, 1978; Hare et al., 2001; Marticorena et al., 2011; Martin & Santos, 2014, 2016; Rothmayr et al., 2011a; Sodian & Thoermer, 2008; Surian et al., 2007). Perhaps as a consequence of this, there currently seems to exists a paradoxical hybrid theoretical framework in the mentalising domain that other-related mental representation is seen as a necessary basis for, and at the same time only exists after, social conflict monitoring (Brass et al., 2009; Deschrijver et al., 2016; Keysar, Lin, & Barr, 2003; Martin & Santos, 2014, 2016; Santiesteban, White, et al.,

2012; Santiesteban, Banissy, et al., 2012; Spengler et al., 2009): The conflict monitoring mechanism assessing action representations in mirror neuron regions is seen in part as helping to define whether the observed body part is *one's own* (also referred to as self-other distinction), which is thought to be a relatively low-level process that is also overarching and assessing conflict between own and other-related mental states represented in the brain. This in turn is needed to support the most high-level ability for human social cognition: Representing mental states of others (Decety & Lamm, 2007; Sowden & Shah, 2014; Spengler, Bird, et al., 2010; Spengler, von Cramon, & Brass, 2010). This interpretation is in part reviewed by many authors by now (Banissy & Ward, 2013; Cook, 2014; de Guzman et al., 2016; Santiesteban, White, et al., 2012; Santiesteban, Banissy, et al., 2012; Sowden & Catmur, 2013; Sowden & Shah, 2014; Sowden, Wright, et al., 2015). Both views on mental representation of others seem extensively backed by empirical (Hare et al., 2001; Keysar et al., 2003; Marticorena et al., 2011; Sodian & Thoermer, 2008; Surian et al., 2007) and theoretical (Brass et al., 2009; Dennett, 1978; Martin & Santos, 2014, 2016; Spengler et al., 2009) arguments. However, the initial development of the false belief task occurred well before a common conflict monitoring mechanism was hypothesized to exist. We thus decided to recourse to philosophy and the debate over the fundamentals of Theory of Mind in this light.

At the basis of the representational mentalising framework lies Dennett's historical argument (1978) that only evidence for understanding another's mental state in a false belief task (i.e., when it is different from your own) can yield definite evidence for a person having mental representation abilities – merely predicting another's behaviour in a true belief task cannot. He added to this that if someone does not adequately responds to a false belief condition "the hypothesis that they impute beliefs and desires to (an)other (individual) would

be dealt a severe blow". Consequently, when an individual fails to pass a false belief task, this can be taken by scholars as evidence that they do not possess the ability to represent others' mental state (Hare et al., 2001; Marticorena et al., 2011; Martin & Santos, 2014, 2016; Priewasser et al., 2018; Ruffman, 1996; Sodian & Thoermer, 2008; Surian et al., 2007). Yet from this initial argument, should this latter conclusion necessarily be true? Note the following example: One can say that if I am employed at a university, that is sufficient to conclude that I have a job. Yet, if I'm not employed at a university, does that mean that I don't have a job? In this example, it is quite obvious that this isn't necessarily the case. Hence, if passing a false belief task guarantees one to have the ability for mental representation, failing a false belief task *could* indicate that I don't have this ability, but it doesn't do so necessarily - especially if a plausible alternative exists for why I may fail a false belief task even while having this ability. Put simply, passing a false belief task indeed suggests an individual's mental representation abilities to be *present*, but failing a false belief task doesn't yield a decisive answer on whether those abilities are *absent*.

Observing a child verbalising their own belief when asked to focus on what the other person thinks is a phenomenon so striking that it is understandably hard for us to imagine anything other than a lack of mental representations having caused it. If a (shared) mechanism exists in the brain that monitors social conflict *after* the other is represented, that modulates the relative expression of own and other-related representations, there may however be an alternative. Such a mechanism being less active could affect dependent measures dependent on their focus: For instance, when one intends to express an own (action) representation (i.e., perform an action), an incompatible other-related action representation would be expected to be inadequately suppressed and thus expressed *too strongly*. This can lead one to execute the other-related action representation instead (or to execute the intended action more slowly;

Brass et al., 2009; Sowden, Koehne, Catmur, Dziobek, & Bird, 2015; Spengler, Bird, et al., 2010; Spengler et al., 2009). In the false belief task, in contrast, the focus of the verbalisation measure lies on expressing the *other-related* (mental) representation (i.e., verbalising Sally's mental state). Remarkably, a lesser active mental conflict monitoring system could yield the exact same results as lacking other-related mental representation: Ineffectively suppressing *one's own* misaligning mental representation to an appropriate extent could interfere with the expression of the other's mental representation, even if adequately *representing* the other's mental state. One may not respond or verbalise the own mental representation instead (see Keysar et al., 2003 for a similar finding of neurotypical adults making errors even while representing others). In sum, observing a child (clinical population, or non-human primate) atypically performing in a false belief situation shouldn't necessarily signify lacking other-related mental representation abilities. See tables 2 and 3 for a summary of the predictions under each framework.

Importantly, if participants may fail a false belief task even while possessing mental state representation abilities, the notion that the human brain does not use other-related mental representation to understand true beliefs (e.g., Martin & Santos, 2016) becomes harder to substantiate: It *could* in principle still be true (the existence of such an alternative interpretation would not present evidence *against* this idea), but it would not be evidenced by the observation that certain populations can pass true belief tasks even while performing atypically in false belief ones, as the latter would then not indicate lacking mental representation abilities necessarily (Hare et al., 2001; Marticorena et al., 2011; Martin & Santos, 2016; Priewasser et al., 2018; Ruffman, 1996; Sodian & Thoermer, 2008; Surian et al., 2007). It is thus still an empirical question whether true belief conditions used in false belief tasks involve the same representational processes as false belief conditions or not. As

a consequence, it may be the case that other-related mental states are represented regardless of the presence of any conflict with own mental states, or of the presence of conflict between the own and the other's perceptual history. Hitherto, scholars haven't conceptually differentiated much between the ideas of *perceptual* conflict detection potentially occurring *before* other-related mental representation and *mental* conflict monitoring occurring *after* both own- and other-related mental states are represented. This presumably resulted in the paradoxical hybrid framework sometimes implicit to the mentalising domain. Yet, even if detecting perceptual conflict detection may help to shape in the observer a neural representation of the other's mental state (regardless of whether it is *required* for it), the potential for mental conflict monitoring occurring *after* they are represented warrants exploring in its own right, as its effects are not accounted for in a representational framework.

Hence, we propose a *reframing* of mentalising data in terms of mental conflict monitoring: Effects derived from a socially incongruent false belief condition could primarily yield insights in extent to which the human brain is able to deal with mental conflict, that is, on the level of mental conflict monitoring rather than on the level of representation (coinciding with perceptual conflict). In sum, we argue for a relational interpretation of mentalising data, rather than the representational one that is currently typical for the field. In this view, a shared social-conflict monitoring mechanism *after* another's mental state (or other social information) is represented could be seen as the brain's most high-level social-cognitive ability, instead of inferring another's mental state or 'Theory of Mind' as such. Authors that have been working on the common conflict monitoring framework have hitherto not gone as far as to reframe data in the false belief task (and in other mentalising tasks) in terms of relational mentalising with implications for how we understand autism and the neural substrates of social cognition, or to argue that the false versus true belief fMRI contrast may

potentially be suited for isolating mental conflict from mental representation. It is these broader implications that we aim to develop in the current paper.

Reconceptualising the Role of Temporoparietal Junction Areas in the False Belief Design

According to meta-analyses, activation in bilateral TPJ and medial prefrontal cortex (mPFC) areas is most often observed in imaging studies of Theory of Mind, both in studies that used a true belief control condition and in those that did not (Lombardo et al., 2007, 2011; Van Overwalle, 2009; Schurz, Aichhorn, Martin, & Perner, 2013; Schurz et al., 2014; Schurz, Tholen, Perner, Mars, & Sallet, 2017). Particularly with respect to the TPJ, influential studies and meta-analyses in the Theory of Mind domain and beyond have strongly favoured a conceptual interpretation of the area as *representing* someone else's mental state (Lombardo et al., 2007, 2011; Overwalle, 2009; Perner et al., 2007; Samson, 2009; Samson, Apperly, Kathirgamanathan, & Humphreys, 2005; Saxe & Kanwisher, 2003; Saxe & Powell, 2006). This has resulted in the hybrid framework sometimes implicit to the domain described above, where a relatively 'low-level' self-other distinction mechanism localised in TPJ subserves a more complex 'high-level' *representation* of belief states within the Theory of Mind domain localised in the very same area (Brass et al., 2009; Lombardo et al., 2007, 2011; Van Overwalle, 2009; Samson, Apperly, Chiavarino, & Humphreys, 2004; Saxe & Powell, 2006; Sowden & Shah, 2014; Spengler et al., 2010). We instead argue that there are methodological reasons to assume that TPJ activity is potentially tied specifically to conflict monitoring even in the mentalising domain.

We performed a systematic review that aims to identify all neuroimaging studies involving a false belief condition, and categorised those on the basis of the type of control

condition used (see Table 4 for the different types of control conditions). The majority of false belief studies did not use a true belief condition as a control (35 out of 51 or 68,6%; listed in the Supplementary Materials). Of those, the majority (19 out of 51 studies or 37.2%) chose a so-called 'false photograph' condition as a control (explained below). The other 16 studies that did not implement a true belief control condition (31.4%) implemented for instance conditions with stories that merely described physical states of objects, which did not involve any mental states or any obvious (perceptual) conflict. This seems to be the result from the main interest within the Theory of Mind domain to focus on the human ability to *represent* others' mental states: Under the representational mentalising framework, the aim is to isolate other-related mental representation from perceptual conflict. Efforts were made to specifically control for perceptual conflict through the use of a non-mental analogue with *physical conflict* (Perner & Leekam, 2008; Saxe & Kanwisher, 2003; Zaitchik, 1990), which involves two representations that are non-mentalistic. In neuroimaging studies using vignettes, it most often consists of the false photograph condition: This could read "A photograph was taken of an apple hanging on a tree branch. The film took half an hour to develop. In the meantime, a strong wind blew the apple to the ground." Here, a difference exists between the state of the world as presented on the photograph, and the state of the world thereafter. The false photograph condition was thought to be structurally equivalent to the false belief condition, including the presence of conflict, except for the non-mental character of the photographs: There exists a visual mismatch between the entities presented in the photograph and the same entities in reality (i.e., the apple in the photograph is hanging on a tree branch whereas in reality it is on the ground). If the methodological aim is to control for (some) perceptual conflict, this may thus prove an appropriate control condition. The conclusion from these studies was that the TPJ is involved in mental representation rather than in social conflict detection.

When the false photograph control conditions were designed (e.g., Perner et al., 2010; Saxe & Kanwisher, 2003; Zaitchik, 1990), the first arguments hadn't been published yet for the potential occurrence of a common social conflict monitoring mechanism (Brass et al., 2009; Spengler et al., 2009), from which one can derive that mental conflict monitoring may occur *after* the own and the misaligning other's mental state are represented in the brain. In a relational framework, a condition aiming to control for all processes *but* other-related mental representation, should evoke in the brain of the observer an own mental representation, and conflict between the own- and an other-related mental representation, but without the other agent evoking in the observer an other-related mental representation as such (see Figure 1, right). Developing such a control is difficult for obvious reasons: How does one generate the specific conflict between the own- and an other-related mental representation without involving an other-related mental state representation? It may seem unreasonable to insist on such a strict control condition, yet without it, one cannot conclude with certainty that results in the false belief condition reflect other-related mental representation only. As far as we are aware, there is no theoretical or empirical reason to assume that the perceptual conflict in a physical conflict condition is identical to or empirically associated with mental conflict (some scholars even argue against this, Perner & Leekam, 2008), meaning that such a control condition may not control for mental conflict. Mental conflict could perhaps even be thought of as non-perceptual, since it most often involves contrasting belief representations about something that is not visible (e.g., an object's *hidden* location). The false photograph condition typically does not involve an own mental state either (it involves only a participant's visually perceiving something), let alone in a balanced way that controls for its particular content. Therefore, with the use of such a control, one also cannot disentangle processes related to *own* mental representation from what is eventually the primary interest under a representational

framework: *other-related* mental representation. When the main aim consists of isolating mental conflict from mental representations, however, a methodologically strict control condition can be more easily developed, as we will explain in the following paragraph. It is for this reason amongst others that we will argue the false belief design to be better suited for making claims about mental conflict monitoring abilities.

If true beliefs engage the same representational processes as false beliefs, neuroimaging experiments should arguably isolate conflict-monitoring processes when they contrast activation in the false belief condition to a balanced true belief condition). Like in the action conflict monitoring literature, the socially incongruent versus congruent contrast should be thought of as revealing processes specifically related to social conflict, if representational processes are filtered out of the contrast. In Table 5, we present the remaining 16 studies of our systematic review, which have contrasted a socially incongruent (false) belief condition against a socially congruent (true) belief condition, because this is the only contrast that could potentially isolate mental conflict monitoring processes from mental representation. Two of those studies definitely did not use a balanced design. Though not always readily stated withing the manuscript, the remaining 14 or 27.5% of all false belief neuroimaging studies may have used a balanced design. Using a balanced design is crucial for assuring that the neural activity produced by the contrast is not influenced by the actual *content* of the representations: Similarly, when isolating neural activity tied to action conflict in two incongruent conditions involving hand actions (e.g., I perform hand action A while observing hand action B; or, I perform hand action B while observing hand action A; Brass et al., 2000, 2005), one would want in the two congruent conditions to consistently use these same hand actions (e.g., we both perform hand action A; or, we both perform hand action B), rather than for instance actions with other body parts (e.g., we both perform feet action A; or, we both

perform feet action B). All but three of the 14 studies that presumably did use a balanced false versus true belief comparison reported TPJ activity. While it is not uncommon to use a region-of-interest (ROI) analysis to identify TPJ activity (e.g., Kovács, Kühn, Gergely, Csibra, & Brass, 2014; Nijhof et al., 2018), it might be the case that with the use of a ROI analysis, TPJ activity would have been detected in some of the studies that did not report such activity. In this light, these data suggest that within the false belief domain TPJ presumably plays a role in mental conflict monitoring. Such an interpretation of TPJ activity in false belief tasks was hypothesized already in the past, even before the true belief condition gained entry as a control in the field (e.g., Mitchell, 2009), but this was not based on the explicit methodological argument we make here, and has so far not found general acceptance within the Theory of Mind domain. It should be noted here that the neuroimaging contrast between false and true belief conditions cannot distinguish between *monitoring* a degree of conflict (e.g., detecting the amount of conflict) and *dealing* with that conflict (e.g., modulating the expression of own versus other-related representations). Table 4 presents a summary of the processes that control conditions are thought to isolate under a representational versus relational framework.

It is important to dwell on our methodological arguments, because it means that the studies that have been most influential in the Theory of Mind domain for the interpretation of TPJ-activity specifically in terms of other-related mental representation only (e.g., Saxe & Kanwisher, 2003; Saxe, Schulz, & Jiang, 2006; Saxe & Wexler, 2005a) may not have been able to cancel out mental conflict monitoring processes (occurring *after* other-related mental representation) from their contrasts. A minority of neuroimaging studies presumably included a *balanced* false versus true belief design (14 in total, or 27.5% of all false belief fMRI studies), which we argue may control for all processes related to mental representation per se. Of those, most reported TPJ activity (Boccadoro et al., 2019; Cracco et al., 2020; Döhnel et

al., 2017; Kovács et al., 2014; Nijhof et al., 2018; Özdem, Brass, Schippers, Van der Cruyssen, & Van Overwalle, 2019; Özdem, Brass, Van der Cruyssen, & Van Overwalle, 2017; Rothmayr et al., 2011b; Sommer et al., 2007, 2018; Wysocka et al., 2020). The implication is that activity in TPJ-areas within false belief conditions may be tied to a social *conflict* processing – even for experiments where a balanced true belief condition is not present. Neural activations found in the influential studies and their follow-ups that used other controls (37 in total, or 72.5%) likely reflected a combination of areas involved in (own and other-related) mental representation per se and in mental conflict monitoring, making it difficult to draw strong conclusions about differential neural localisation of these processes from these experiments alone. Yet, we think that no false belief study today has used a control condition that can guarantee the involvement of the TPJ in other-related mental state representation. Overall, if it can be confirmed that true beliefs involve other-related representation, like would be expected under a common conflict monitoring framework for human social cognition, it will be important in the future to be consistent in the use of a balanced socially congruent control condition when interested in mental conflict monitoring, while the use of physical conflict controls to isolate other-related mental representation may need to be discontinued.

Relational Mentalising: Mirroring Mental States?

In the previous Section, we argued that the false versus true belief contrast of the Theory of Mind domain may implicate TPJ-activity in mental conflict more so than mental representation. If this is the case, where does the brain represent other-related mental states? And are these represented in the same brain areas responsible for *own* mental representations, i.e., do we *mirror* mental states (Van Overwalle, 2009)? Which mentalising designs do we currently have that could yield answers to these questions?

In the action perception domain, other-related representations are primarily investigated by comparing an other-representation only condition against a control condition that has no own- or other-related representations (see Figure 1A). The passive observation of another's actions should make sure that the motor cortex contains an other-related action representation, but *no* own action representation, as the participant is not actively moving. This ensures that no action conflict is present in the action observation task, as there is no own-related representation in the premotor cortex for the other-related representation to diverge from. Similarly, in order to focus on other-related mental representations only in the mentalising domain, one could look for a design that implements other-related mental representations, but *no* own mental representations. Such tasks include, for instance, the Movie for the Assessment of Social Cognition (MASC; a 15-min video about 4 characters getting together for a dinner party about which the participant answers questions concerning the characters' feelings, thoughts and intentions; Dziobek et al., 2006), and the Frith-Happé Animations Task (which presents participants with simple geometric shapes that are moving in a way that evokes a sense of intentions; Castelli et al., 2000). These tasks might appear to isolate mental representation, analogous to action observation tasks, because they do not contain obvious mental conflict in the way that false belief tasks do. The finding that these tasks typically evoke TPJ-activation therefore seem to support the role of this area in mental representation.

It is useful to keep in mind the main limitation that distinguishes the mentalising domain from the action perception domain. Human thoughts are always unconstrained: Participants can't be made to keep their mental states completely 'absent' in their brain as easily as they can be asked to perform no motor actions. However, 'absent' own mental states

are crucial for the validity of conclusions that these designs yielding insights in mental representation only. If there are unconstrained own mental states, they could well be in conflict with the other-related mental representation that one does not necessarily share oneself (i.e., the participants themselves are not experiencing the dinner party). This type of *divergence* between mental states, without *directly conflicting beliefs* like which occurs in false belief tasks, can be termed 'latent mental conflict'. In effect, the occurrence of latent mental conflict could result in such mentalising tasks boiling down to being one type of social congruency design, even if this was undesigned: Its dependent measures may thus capture latent mental conflict monitoring processes in addition to mental representation. Like in the false belief domain, neuroimaging or behavioural measures that assess understanding of the other's mental state used in the MASC or the Frith-Happé Animations Task (e.g., belief verbalisations) or any other task that uses a similar design, could be seen as dependent measures reflecting (latent) conflict monitoring (the latter with a focus on expressing the other-related representation; see Table 3). Even while latent mental conflict may occur in the classical Sally-Anne task as well, the false versus true belief contrast should cancel out this conflict as unconstrained own mental representations (next to the manipulated own belief) and thus latent conflict with the manipulated other-related belief could presumably be present in both conditions, making it the most purely controlled design at present for isolating mental conflict from mental representation.

One other reason why the TPJ has been particularly tied to mental representation rather than to mental conflict monitoring, is that a specific vignette study that implemented story lines that describe actions of what they called a character's 'true beliefs' found that the TPJ-response was also evident in these conditions, which was taken as suggesting that the area is involved in mental representation per se (Saxe & Powell, 2006). Such a specific 'true belief'

vignette would for instance read “Rob tied his dog’s leash to a lamppost while he went into a store to buy coffee. When he came out, his dog had run across the street. He guessed that the leash had come untied.” Even while the character’s eventual belief can be considered as true, the state of the world as understood by the participant is shifting from the character’s mental state (here: knowing that the dog is running across the street in spite of knowledge of Rob’s belief that the dog’s leash is tied). Thus, we want to point out that this particular ‘true belief’ condition, in contrast with a true belief condition of classical Sally-Anne tasks where the state of the world does not change (i.e., the object in the Sally-Anne task is not relocated), do raise the spectre of the other being misinformed with respect to what the participant knows is true for at least for a brief period in time (i.e., before the storyline concludes that Rob updates his false belief in order for it to become ‘true’). Hence in principle it could also be mental conflict that elicits TPJ activity in this specific ‘true belief’ condition. The key difference that should lead to TPJ activity, in our framework, is not whether the other’s mental state is eventually ‘true’, but rather whether a mismatch is present in the design between another’s mental state and the participant’s own understanding of the world. Also in these studies, there is no immediate presence of a condition that specifically focusses on controlling for (latent) mental conflict.

Especially given the hypothesised existence of ‘latent mental conflict’, a consequence of the methodological limitation that the mentalising domain is subject to, it is difficult to draw strong conclusions on the nature of neural representations of own and others’ mental states per se on the basis of false belief data and data from other present mentalising data. The possibility that the TPJ is tied to mental conflict in tasks of other mentalising designs should be confirmed in future research that tries to disentangle ‘latent’ mental conflict and other-related representations from one another in these specific designs. Without it, the question

whether other-related mental states are indeed 'mirrored' in brain areas that also represent own mental states remains unanswered.

Implicit Theory of Mind as a Measure of Conflict: Valid or Not?

In the action perception domain, behavioural effects obtained in socially incongruent versus congruent conditions are interpreted in relational terms: For example, larger RTs in the incongruent (versus the congruent) condition are described in terms of the extent to which the individual processes the conflict inherent to the former condition (and manages to suppress the other-related representation), rather than in terms of whether the individual represents the action. In the Theory of Mind domain, however, behavioural measures (such as interaction behaviour; Buttelmann, Carpenter, & Tomasello, 2009; looking times; Gliga et al., 2014; and RTs; Kovács et al., 2010), are often attributed to representation of others' mental states. In particular, such dependent measures are often considered to yield insights into more implicit forms of mental representation termed 'Implicit Theory of Mind'. Tasks in this domain use stimuli with manipulations based on the false beliefs in the Sally-Anne task, but typically use dependent measures that do not focus on individuals showing explicit understanding of the other's mental state (e.g., in belief verbalisations). The manipulations are considered implicit both because of their task-irrelevancy as well as the fact that participants are usually unaware of the belief manipulations and are thought not to use linguistic deliberation regarding them (Deschrijver et al., 2016; Schneider, Bayliss, Becker, & Dux, 2012; Schneider, Lam, Bayliss, & Dux, 2012; Schneider, Nott, & Dux, 2014; Schneider, Slaughter, Becker, & Dux, 2014b). The distinction between explicit versus implicit Theory of Mind has gained momentum over the past 10 years, especially after findings that certain populations perform consistently differently in the two types of tasks. A number of different 'two-systems' accounts of mentalising have been put forward (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013;

Carruthers, 2016). Such accounts have primarily proposed an early-developing and more or less automatic form of implicit mental representation, in addition to explicit mental representation abilities that are more slowly developing while being cognitively demanding (Apperly & Butterfill, 2009; Butterfill & Apperly, 2013; Carruthers, 2016, 2017).

We suggest however that under a relational framework, implicit as well as explicit measures in false belief tasks can be reframed as one mechanism of mental conflict monitoring rather than 2 types of mental representation (see Tables 2 and 3). Specifically, dependent measures with no inherent focus on the content of represented mental states (e.g., some eyetracking and neuroimaging measures) may be more directly driven by individuals' processing of mental conflict, regardless of the false belief task being implicit or explicit. In the same vein, the action conflict domain considers TPJ activity found in an incongruent versus congruent action condition contrast as yielding direct insights in action conflict monitoring processes. Consistent with this, imaging results in false belief tasks generally show the involvement of the TPJ in both lines of research, also when false versus true belief comparisons are used (Bardi et al., 2017; Boccadoro et al., 2019; Cracco et al., 2020; Dodell-Feder et al., 2011; Hyde et al., 2015; Kovács et al., 2014; Overwalle, 2009; Saxe & Kanwisher, 2003; Saxe, Moran, Scholz, & Gabrieli, 2006; but see Naughtin et al., 2017b). Verbal measures or 'explicit' mentalising (e.g., Baron-Cohen et al., 1985), instead, tap more strongly into the expression of *other-related* representations (e.g., when answering the question 'Where will Sally look for the ball?'). Few dependent measures in implicit false belief tasks, in contrast, have such a focus: Precisely because of their aim of assessing *spontaneous* mentalising processes, they often do not inherently focus on any mental representation in particular (e.g., eyetracking measures). Reaction time measures expressing a violation of the

participant's own belief in an implicit false belief task may rather be tapping more strongly into expressing the *own* representation (see further in this Section; Bardi et al., 2016; Deschrijver et al., 2016; Kovács et al., 2010; Martin & Santos, 2014), just like explicit measures that focus on expressing the own belief (e.g., verbalising the own mental state, Apperly et al., 2004; Samson et al., 2004; Sommer et al., 2018). In the next Section, we will show that results in the developmental and clinical trajectory of these tasks may align in more consistent patterns when taking into account how mental conflict monitoring may play out differently depending on the focus of the dependent measure (see Tables 2 and 3 for a summary), rather than when relying on the current categorisation in terms of implicit versus explicit Theory of Mind.

As an example of the two types of dependent measures within the implicit Theory of Mind domain (mostly those with a relatively strong focus on the expression of the own mental representation and those with no focus on any mental representation in particular), one of the most known tasks in this domain used in adults consists of a design where participants are asked to detect the presence of a ball while they themselves as well as an agent hold a belief about whether or not the ball will be present (Kovács et al., 2010). Specifically, participants observe movies in which an agent forms a belief about the location of a ball, which can either be behind an occluder or roll out of the scene. The agent walks out of the scene, and while he is away the participant also forms a belief about the ball's location. After the agent walks back in, the occluder falls down, and participants have to press a button when the ball is present. Whether the ball is behind the occluder is however random, and independent of what happens during the movie. In general, participants are expected to be slower to detect the ball when they had believed the ball not to be there, as compared to when they had believed the ball to be present. In this sense, scholars have asserted that it may in the first place reflect a violation

of the *own* belief regarding the ball's location (Bardi et al., 2016; Martin & Santos, 2014), meaning that the reaction measure would have a relatively strong focus on the expression of the *own* mental representation. It was shown that when the participant believed that the ball would not be present, they were faster at detecting the presence of the ball if the agent held the belief that the ball *would* be present (i.e., a socially incongruent condition), compared to a where both thought that the ball would not be present (i.e., a socially congruent condition). It was thought that the belief of the agent aided the participant in detecting the ball, by speeding up their RTs when the agent believed the ball would be present (Deschrijver, Bardi, Wiersema, & Brass, 2016; Schneider, Lam, et al., 2012; Schneider, Nott, et al., 2014; Schneider, Slaughter, et al., 2014). Many other implicit Theory of Mind studies use dependent measures without a particular focus: When performing the same task, 7-month-olds looked longer at the absence of the ball when only the agent had believed the ball to be there (i.e., socially incongruent), as compared to when both believed the ball would not be there (i.e., socially congruent). In principle, the difference in the conditions of interest entails a difference in alignment between the beliefs of the agent and of the participant (i.e., mental conflict), and thus the results may more directly signify the participants processing of mental *conflict* rather than representation per se; see also Deschrijver et al., 2016).

It should be noted that the field of implicit Theory of Mind currently is controversial, as a debate is going on with respect to the validity of many of its results. For example, it has been argued that the reaction time results in the task mentioned above may be generated by timing differences between conditions for an attention check in the original task (a button press required from the participant at the moment where the agent leaves the scene, Phillips et al., 2015). Other scholars have remarked that the results in implicit false belief tasks may be generated by so-called 'submentalising' processes, which may result from domain-general

cognitive processes which simulate the effects of mentalising (Heyes, 2014), for instance 'attention-grabbing' differences between conditions. If these criticisms are borne out, it would suggest that the results of these implicit false belief tasks should be interpreted neither in terms of mental representation nor in terms of conflict monitoring. In addition, there have recently been some large-scale studies that variously reported successful, partial and non-replications of specifically anticipatory looking time results, indicating that more research is required to establish the robustness of these results (e.g., Kulke, Duhn, Schneider, & Rakoczy, 2018; Kulke, Johannsen, & Rakoczy, 2019; Kulke & Göttingen, 2017). However, arguments against submentalising interpretations of implicit Theory of Mind data include the involvement of core mentalising regions such as the TPJ in implicit tasks (Bardi et al., 2016; Bardi et al., 2018; Bowman, 2015; Filmer et al., 2019; Hyde et al., 2015; Kovács et al., 2014; Naughtin et al., 2017; Nijhof et al., 2016; Nijhof et al., 2018; Schneider, Slaughter, et al., 2014), a relationship of results with traits of autism (Deschrijver et al., 2016; Nijhof, Brass, & Wiersema, 2017) and similar results to the original implicit mentalising task described above in a recent study which removed the timing differences between conditions from the task (El Kaddouri, Bardi, De Bremaeker, Brass, & Wiersema, 2019; for still other arguments, see Schneider, Slaughter, & Dux, 2017).

In sum, there is ongoing debate about the validity of results within the implicit Theory of Mind domain (Heyes, 2014; Kulke et al., 2018, 2019; Kulke & Göttingen, 2017; Phillips et al., 2015), though not all implicit Theory of Mind findings have been challenged (e.g., looking time measures other than anticipatory looking times). How this debate resolves will have implications for whether the existing implicit false belief tasks can be treated as measures of mental conflict processing or not. Overall, however, understanding results in studies of false belief tasks while appreciating the focus of their dependent measures, more so than their

presumed reliance on explicit versus implicit processes, may be useful going forward.

Others' Representations *Diverging From Ours:* At the Heart of Social Cognition?

Over the last 40 years, the false belief design has helped to shape the idea that humans should foremost be able to infer others' mental states or 'mindread' in order to achieve social success (Baron-Cohen et al., 1985). This focus on mental representation has led to the widespread assumption that an *atypical* effect in false belief conditions signifies lacking mental *representations*. However, following the arguments presented in the previous Section, an atypical effect in a false (versus true belief) condition could instead reflect an inactive mental conflict monitoring system, with mental representations per se intact. This is particularly important for interpreting data focusing on autism, where atypical data in false belief conditions have consistently been presented as evidence for lacking mental representations or *mindblindness* (e.g., Senju, Southgate, White, & Frith, 2009). We acknowledge that scholars (e.g., de Guzman et al., 2016; Sowden & Shah, 2014) have previously suggested that conflict monitoring may contribute to differing responses in mentalising tasks in autism (as proposed by Spengler, Bird, et al., 2010 and also further discussed in a developmental context), but together with the common assumption that false belief tasks predominantly reflect mental representation, and given the assumptions of the representational mentalising framework (see Table 1), this has resulted in a view where social conflict monitoring differences are seen as a potential origin of mindblindness in autism. In contrast, we propose that individuals on the autism spectrum are not 'mindblind', and discuss how result patterns across studies with distinct dependent measures provide evidence that social difficulties in individuals on the autism spectrum may result from differences in mental conflict monitoring. We outline a similar argument with respect to lacking effects in

developmental and non-human primate data: If results from implicit false belief data can indeed be taken as reflecting mentalising processes, a lack of effects in a socially incongruent condition at very young ages in neurotypical infants, or in non-human primates, can arguably signify mental conflict monitoring difficulties rather than lacking mental representation abilities *per se*.

Autism: An Iconic Clinical Case for Understanding Mentalising

While most neurotypical children pass traditional false belief tasks requiring explicit verbalisation of others' beliefs at the age of four, children on the autism spectrum often perform these tasks more poorly, despite having at least an equivalent mental age (Baron-Cohen et al., 1985). Complicating the evidence for the Theory of Mind Hypothesis of autism, however, is the fact that about one fifth of these young children on the autism spectrum *do* pass false belief tasks, and that older children and adults with high-functioning autism usually display typical behavioural performance as well (Baron-Cohen, Jolliffe, Mortimore, & Robertson, 1997; Peterson & Slaughter, 2007; Scheeren, De Rosnay, Koot, & Begeer, 2013). As a result, tests that assess verbalisation abilities in a Theory of Mind design do not always grasp the actual social difficulties individuals on the autism spectrum face throughout their lives.

Nevertheless, lacking mental representations or 'mindblindness' are commonly thought to be core to the disorder for two reasons: First, it has been suggested that the typical performance of older individuals on the autism spectrum in verbalising other's beliefs may follow from the use of *compensatory strategies* or behavioural rules to complete the task at

hand (Deschrijver, Bardi, Wiersema, & Brass, 2016; Happé, 1995; Schneider, Slaughter, Bayliss, & Dux, 2013; Senju, 2013a; Senju, Southgate, White, & Frith, 2009; Zwickel, White, Coniston, Senju, & Frith, 2011). In other words, cognitive mechanisms involved in representing other people's mental states are still thought to be affected in both children and adults on the autism spectrum, but with the latter more able to compensate for this with the use of explicit reasoning or the like. Second, neuroimaging and eyetracking measures in false belief paradigms have mostly yielded evidence for differences between groups with and without autism, even if there are no differences between groups in their ability to verbally report others' beliefs (Burnside, Wright, & Poulin-dubois, 2017; Gliga et al., 2014; Nijhof et al., 2018; Schneider et al., 2013; Schuwerk, Jarvers, Vuori, & Sodian, 2016; Schuwerk, Vuori, & Sodian, 2015; Senju, 2013a, 2013b; Senju et al., 2009; White, Frith, Rellecke, Al-noor, & Gilbert, 2014; Zwickel et al., 2011). Scholars have also shifted their attention to testing *implicit* false belief paradigms in individuals on the autism spectrum because it was hypothesized that being explicitly prompted to consider another person's belief may draw responses from those participants that are not representative of how they spontaneously *represent* others' mental states. Together with the idea of the representational framework that these dependent measures, including neuroimaging and eyetracking data in false belief tasks, reflect the ability to *represent* mental states, the logical conclusion is that mental representation or 'mindreading' is affected in autism.

A population that is 'mindblind' (Alcalá-López et al., 2019; Baron-Cohen et al., 1985; Lombardo & Baron-Cohen, 2011; Senju et al., 2009), however, should in priniciple be expected to show consistently decimated effects across all dependent measures: A completely lacking other-related mental representation should never have any influence. An inactive mental conflict monitoring system, in contrast, would lead to more subtle differences, as this

would involve other-related mental representation per se to be intact: One could expect consistently diminished effects across dependent measures that may capture the methodological manipulation of mental conflict most purely (e.g., neuroimaging; White, Frith, Rellecke, Al-noor, & Gilbert, 2014; Nijhof et al., 2018; and some implicit dependent measures; Burnside et al., 2017; Gliga et al., 2014; Schneider et al., 2013; Schuwerk et al., 2016, 2015, Senju, 2013a, 2013b; Senju et al., 2009; Zwickel et al., 2011). Dependent measures that focus on expressing own- or other-related mental representations may yield slightly more inconsistent results, since they focus on expressing an *intact* representation in a design that manipulates social conflict. If anything, however, one could expect to find a *lesser*-than-typical influence of the other-related mental representation when trying to expressing this representation (e.g., verbalising the other's belief; Baron-Cohen, Jolliffe, Mortimore, & Robertson, 1997; Peterson & Slaughter, 2007; Scheeren, De Rosnay, Koot, & Begeer, 2013); and a *stronger*-than-typical influence of the other-related representation when expressing the *own* representation (e.g., potentially other implicit dependent measures (Deschrijver et al., 2016), or verbalisation of the own mental state; e.g., Apperly et al., 2004; Samson et al., 2004; Sommer et al., 2018). This would signify, respectively, an ineffective suppression of the participant's *own* mental representation and of the *other-related* mental representation. In particular, this latter result would never be expected if other-related mental representations were lacking, as these representations need to be present before they can hinder performance.

Neuroimaging studies of mentalising in autism spectrum disorder are especially interesting in this respect as they may yield a clearer view on mental conflict monitoring through the specific use of the false versus true belief contrast (if the latter condition involves the same representational processes as the former). Compared to matched control groups, studies have reported decreased activity in the TPJ for false belief (versus true-belief)

conditions in populations on the autism spectrum versus controls (White, Frith, Rellecke, Al-noor, & Gilbert, 2014; Nijhof et al., 2018), though some studies have reported similar or increased TPJ activity in autistic populations (Dufour et al., 2013; Sommer et al., 2018). Findings of lower TPJ activity in autism (as compared to either true belief conditions or non-mental analogues) have however consistently been reported in other types of mentalising designs as well (which we assert, may reflect 'latent mental conflict'.; Kana et al., 2015; Kana, Keller, Cherkassky, Minshew, & Just, 2009; Kana, Libero, Hu, Deshpande, & Colburn, 2014; Koster-hale, Saxe, Dungan, & Young, 2013; Lombardo et al., 2011; Murdaugh, Nadendla, & Kana, 2014; Spengler et al., 2010; White et al., 2014; but see Mason, Williams, Kana, Minshew, & Just, 2008). Importantly, the reduced TPJ-activity in individuals on the autism spectrum has also been found to correlate with actual social difficulties in everyday life (Kana et al., 2009; Lombardo et al., 2011; White et al., 2014), in contrast to some verbalisation measures of mental representation (Baron-Cohen et al., 1997; Peterson & Slaughter, 2007; Scheeren et al., 2013). Together this may suggest that the processing of mental conflict is affected in autism.

So-called implicit dependent measures often do not focus on any mental representation in particular, meaning that they are more directly dependent on the methodological contrast of the design, which may largely tap into mental conflict processes in the case of a false belief versus true belief contrast (provided that they indeed tap into mentalising processes). Differences in looking times for socially incongruent (versus congruent) mentalising conditions are reduced for individuals with (a higher likelihood of being diagnosed with) autism of a variety of different age groups, as compared to matched controls (Burnside et al., 2017; Gliga et al., 2014; Schneider et al., 2013; Schuwerk et al., 2016, 2015, Senju, 2013a, 2013b; Senju et al., 2009; Zwickel et al., 2011). Thus, while neurotypical adults are processing

that others' mental states conflict with their own, individuals on the autism spectrum may be less so. In other words, the extent to which parts of our visual environment are made more salient to us when we infer other people's mental states is likely to be modulated by whether those mental states conflict with our own or not, and hence the looking behaviour of participants who are processing mental conflict to a lesser extent (here: individuals on the autism spectrum) may be driven more weakly by the conflict in other people's mental states in false belief designs.

In addition, one recent implicit false belief task has yielded more direct evidence for potential mental conflict monitoring difficulties in autism (Deschrijver et al., 2016). This task, described in an earlier Section, was developed to merely require a key response to the detection of a visual target, while using belief manipulations similar to those in the Sally-Anne task. It can be speculated that ball detection RTs may reflect primarily an expression of the participant's *own* belief or expectation about the location of the ball (Bardi et al., 2016; Martin & Santos, 2014). The key finding of this task was that *another* person's belief that the target would be present can *speed up* detection of the target in participants, when the participant themself had been convinced that the target would not be present (Deschrijver et al., 2016; Kovács et al., 2010). In this context, the Theory of Mind Hypothesis of autism, which asserts a lack of belief representations, would predict that adults on the autism spectrum should not experience this benefit to detection performance. Strikingly, however, adults in the clinical group with stronger autism traits were significantly *slowed down* in the detection task. These subjects were thus *hindered* in detecting the visual target when the other person's belief conflicted with their own (Deschrijver et al., 2016), as could be expected when experiencing a lesser active conflict monitoring system in a dependent measure with a focus on expressing the own mental state. In order to be slowed down by another person's conflicting belief, these

individuals must have first represented it. In another explicit study, individuals on the autism spectrum seemed to be hindered by the other's false belief more strongly than controls when they needed to verbalise their own mental state, another dependent measure focussing on expressing the own mental representation (Sommer et al., 2018). Although it should be noted that these are just two studies (Deschrijver et al., 2016; Sommer et al., 2018), and in one study hindrance was only observed in those of the clinical group with the strongest autism traits (Deschrijver et al., 2016), the results cannot be explained by lacking other-related mental representations, and cannot be accounted for altogether under a representational interpretation framework.

This empirical finding shares a parallel with some in the action conflict domain: With autism as an important clinical case, it was at first argued that difficulties to understand others in real life may follow from an inability to *represent* others' actions (i.e., the 'broken' mirror hypothesis of autism; Iacoboni & Dapretto, 2006; Williams, Whiten, & Singh, 2004; see also Figure 1, left). More recently, however, even though no meta-analytic evidence for a group difference was detected, exactly half of autism studies included in a meta-analysis on automatic imitation (5 out of 10) reported a relationship of autism with hyperimitation effects, that is, a larger-than-typical influence of an incompatible observed other-related action on the execution of an own hand action. The focus of the dependent measure on expression of an own (intact) action representation (i.e., executing the own action intention) may have something to do with the inconsistency of the results. Autism studies in action conflict using dependent measures without any particular representational focus (e.g., neuroimaging) are needed to provide a more consistent answer as to whether action conflict monitoring is affected. In any case, the results suggests that neural representation of others' actions is not 'broken' in the autism spectrum, as in order to be sensitive to another person's conflicting

actions at all these must first have been represented. The idea that social cognitive difficulties characteristic of autism reflect an inability to *represent* others' actions has thus fallen out of favour (Fan, Decety, Yang, Liu, & Cheng, 2010; Southgate & Hamilton, 2008; Sowden, Koehne, et al., 2015; Spengler et al., 2010). Similarly, the finding that individuals on the autism spectrum can be hindered more strongly (or to the same extent) by another person's incongruent mental state when expressing an own mental state suggest that neural representation of others' mental states is not 'broken' in autism.

Lastly, participants might give unusual verbal responses when asked to verbally report other people's beliefs if the system monitoring conflict between mental states of the agent and oneself is not active, even if able to flawlessly grasp *what* the other thinks. An interpretation of the verbalisation results in terms of ineffective social conflict monitoring in autism has been presented before in the literature (Bloom & German, 2000; Brass et al., 2009; Ozonoff et al., 1991; Pellicano, 2007; Spengler et al., 2010, 2009; for reviews, see Banissy & Ward, 2013; Cook, 2014; de Guzman, Bird, Banissy, & Catmur, 2016; Santiesteban, White, et al., 2012; Santiesteban, Banissy, Catmur, & Bird, 2012; Sowden & Catmur, 2013; Sowden & Shah, 2014; Sowden, Wright, Banissy, & Catmur, 2015) but these differences have widely been regarded as an underlying reason of why 'mindreading' may be affected in autism. When consistently accepting *the* most seminal argument of the Theory of Mind domain (Dennett, 1978), namely that passing belief verbalisation in a false belief task guarantees in a participant the ability for other-related mental representation, the well-preserved belief verbalisation abilities in adults on the autism spectrum in false belief tasks suggest that their mental representation abilities are intact. Not consistently accepting this argument seems a slippery slope: In that case, one may as well interpret findings where any individual (child, adult, primate, …) shows understanding of a false belief as signifying that they, too, used strategies

instead of mental representation abilities to solve the task.

There are several ways in which an account of mentalising differences in autism in terms of mental conflict monitoring combined with intact other-related mental representation may be more parsimonious than lacking mental representation per se. First, it does not require the supposition that compensatory strategies explain the typical performance of most individuals on the autism spectrum to verbalise others' mental states (Baron-Cohen et al., 1997; Peterson & Slaughter, 2007; Scheeren et al., 2013). Rather, this can simply be accounted for by mental representation skills being intact. Second, impaired mental conflict monitoring can explain why false versus true belief contrasts of fMRI/eyetracking, but not belief verbalisation measures show a relationship with actual social difficulties in autism. The former but not the latter should yield a relatively pure reflection of their (impaired) mental conflict processing abilities. Third, it offers an alternative to the idea of individuals on the autism spectrum lacking 'implicit' but not 'explicit' mental representation abilities (see earlier for an outline on two-system accounts of mental representation; Apperly & Butterfill, 2009; Butterfill & Apperly, 2013; Carruthers, 2016, 2017). Specifically, the performance patterns in those clinical populations can be explained by measures of belief verbalisation being more strongly influenced by intact other-related mental *representation* abilities (as dependent measures that focus on the expression of either own or other-related mental representations). An interpretation in these terms would also clarify why individuals on the autism spectrum show reduced TPJ activity in both implicit and explicit versions of mentalising tasks (e.g., Nijhof et al., 2018; White et al., 2014), in spite of not showing belief verbalisation difficulties in the latter (Baron-Cohen et al., 1997; Peterson & Slaughter, 2007; Scheeren et al., 2013). Because the TPJ is considered a core social brain area and implicit false belief findings in autism have been found independent of performance in executive processing tasks (Schuwerk

et al., 2016), we think it is unlikely that executive rather than social conflict processes are underlying the effects, though attentional differences between the two groups may exist.

Relational Mentalising in Developmental and Nonhuman Primate Populations

In the previous Section, we described how re-assessing data from a socially incongruent design within a relational framework may play out in a population on the autism spectrum. Here, we sketch how a similar analysis may apply to research focussing on false beliefs in developmental and non-human primate populations, where atypical effects in false belief conditions are also commonly seen as indicating a lack of other-related mental representation abilities.

In the early days of developmental mentalising research, the role of conflict in false belief tasks was often touched upon, as it was shown that executive functions such as inhibitory control, which may allow an individual to suppress their own conflicting knowledge of current reality (Pellicano, 2007), may be key to the development of belief verbalisation abilities (Carlson, 2002, 2010; Carlson, Mandell, & Williams, 2004; Carlson & Moses, 2001; Carlson, Moses, & Claxton, 2004; Hughes, 1998). Such studies still focused on mental representation, however, with the monitoring of perceptual conflict seen primarily as needed for mindreading and thus other-related mental representation to occur, partly aided by executive functions. For instance, performance on mentalising tasks developmentally follows pre-schoolers' successful performance on tasks of inhibitory control (Flynn, O'Malley, & Wood, 2004). Individual differences in false belief verbalisation and individual differences in executive functions show robust associations in neurotypical children of various ages (Carlson, 2002, 2010; Carlson, Mandell, et al., 2004; Carlson & Moses, 2001; Carlson, Moses,

et al., 2004; Hughes, 1998). In children on the autism spectrum, on the other hand, performance is significantly worse on both false belief and executive function measures relative to control children, independent of intellectual functioning (Joseph & Tager-Flusberg, 2004; Ozonoff et al., 1991; Zelazo, Jacques, Burack, & Frye, 2002), and a significant correlation was found between measures of executive function and mentalising in a group of children on the autism spectrum (Ozonoff et al., 1991). Nevertheless, false-belief tasks cannot be construed entirely as executive tasks as children on the autism spectrum have been found to pass non-mental analogues of false belief tasks but not false belief tasks themselves, suggesting that it is not the executive functions per se that predict their performance on belief verbalisation measures (Leekam & Perner, 1991; Leslie & Thaiss, 1992; but see Russell, Saltmarsh, & Hill, 1999). Moreover, one study in autism found that while almost the entire group showed impairment in executive functions, only half of them showed concurrent difficulties in belief verbalisation, showing that the relationship between the development of executive functions and that of belief verbalisation is not absolute (Ozonoff et al., 1991). From the perspective of a relational framework, what should be taken away from these studies is that the development of executive functions like cognitive control may in part help to suppress the participant's own mental representation when required to verbalise another person's mental representation, rather than helping to detect perceptual conflict needed for mental representation, as may be assumed under a representational framework. However, executive functions like cognitive control cannot completely explain mental conflict monitoring. A similar idea of executive functions possibly aiding, but not being equivalent to, the mechanism that monitors conflict between own and other-related representations exists in the action conflict domain as well (Cracco et al., 2018; Cross, Torrisi, Reynolds Losin, & Iacoboni, 2013; but see Brass et al., 2005).

In order to investigate Theory of Mind at very young ages, imaging and behavioural measures (including looking times) are used to assess mentalising abilities independent of children's verbalisation abilities. Interpretations of such developmental data have mainly been presented in terms of whether or not very young children *represent* mental states. Here, again, performance is often compared between socially incongruent and congruent mentalising conditions, meaning that neuroimaging and behavioural data could be indicative of mental conflict monitoring processes more so than mental representation per se. Neurotypical infants of 2 years or below are found to distinguish between incongruent versus congruent trials in a number of mentalising studies, even without mastering verbalisations of beliefs. This was shown, for example, by longer looking times when an actor's searching behaviour is incongruent with the location of a toy (Onishi & Baillargeon, 2005; Senju et al., 2009; Surian et al., 2007; Wiesmann, Friederici, Singer, & Steinbeis, 2017). One study even reported results in 7-month-olds (Kovács et al., 2010): As a sign of a violation of expectation, infants show longer looking times to a condition where an agent had believed the ball would be present and the infant had not (a socially incongruent condition) as compared to when both the infant and the agent had believed that the ball would not be present (a socially congruent condition). Children with a higher likelihood of developing autism, instead, show diminished looking differences for incongruent versus congruent mentalising conditions at young ages (Burnside et al., 2017; Gliga et al., 2014; Schuwerk et al., 2016). From all this, authors have concluded that 'implicit belief *representation'* may develop within the first year of life (Baillargeon, Scott, & He, 2010; Alan Leslie, Friedman, & German, 2004; Scott & Baillargeon, 2017; Scott, Baillargeon, Song, & Leslie, 2010; Wang & Leslie, 2016) or even be present from birth throughout life (Kovács et al., 2010; Schneider et al., 2017), while being affected from early ages on in autism.

Re-evaluating false belief tasks as social congruency designs suggests that neurotypical individuals may possess a mechanism for monitoring conflict between an own and an other-related mental representation from a very young age, and into adulthood, while this mechanism may be affected in autism (Grainger, Henry, Naughtin, Comino, & Dux, 2018). Such a reframing of results is also important to understand studies which reported lacking effects in incongruent mentalising conditions for young infants: Methodologically speaking, this could be explained by reduced processing of mental conflict, not an absence of mental representations necessarily. Even with intact representations one may not monitor the conflict between own- and other-related representations very well. While there are over 30 original research papers concluding that mentalising abilities exist at early ages (as stated in Scott & Baillargeon, 2017), some studies have failed to replicate original findings, reporting null results both in children, adults, and elderly adults, although only in certain specific dependent measures such as anticipatory looking (Burnside et al., 2017; Kulke et al., 2018, 2019; Kulke & Göttingen, 2017; Powell, Hobbs, Bardis, Carey, & Saxe, 2018). The analysis of mentalising data presented here should be regarded in light of how the ongoing debate about how well implicit false belief tasks measure mentalising processes (Heyes, 2014; Phillips et al., 2015) further develops.

Similar remarks go for the study of mentalising in non-human primates. Scholars have typically sought to explain differences in performance between false belief tasks and true belief tasks as an indication of social cognition in non-human primates relying on processes other than mental representational ones (Hare, Call, & Tomasello, 2001; Marticorena, Ruiz, Mukerji, Goddu, & Santos, 2011; for a summary, see Martin & Santos, 2014, 2016). The most influential theoretical accounts have, for instance, focussed on these animals' use of abstract behavioural rules, so-called minimal Theory of Mind, awareness relations, etc. (Martin &

Santos, 2014, 2016; Phillips & Norby, 2019). The sense that primates that failed a false belief task likely did not have mental representational abilities, resulted in a need to explain what other processes non-human primates might be using to perform in true belief conditions. However, in light of evidence for a common conflict monitoring mechanism in human social cognition (Brass et al., 2009; Santiesteban, White, et al., 2012; Santiesteban et al., 2015; Santiesteban, Banissy, et al., 2012; Spengler et al., 2009), failing a false belief task shouldn't necessarily be taken as indicating a lack of mental representations in non-human primates either. Theoretical accounts of primates' social cognition may thus need to be considered in light of the role mental conflict monitoring may play in false belief tasks.

All in all, both developmental and non-human primate research may benefit from a relational framework, and from taking into account the patterns of results in relation to the specific focus of the dependent measures used. This may include focussing research on the designs and dependent measures that are most optimal for distinguishing the hypotheses of lacking mental representations versus ineffective mental conflict monitoring (see Table 3).

Representational versus Relational Interpretations in the Larger Social-cognitive Domain.

Over different social-cognitive fields, social conflict conditions have been interpreted in distinct ways – sometimes in terms of representational processes (e.g., the Theory of Mind domain, but also the perspective taking domain, and that of lying, moral dilemmas, irony, sarcasm and humour) and sometimes in terms of relational processes (e.g. the action mirror neuron domain, but also the touch perception and empathy domain). Perhaps not coincidentally, the domains that utilise a representational framework are those experiencing

the methodological limitation mentalising researchers have always faced, that is, that the participant's own representations can hardly be ruled out. For example, it is hardly possible to prevent a participant of having their own visual perspective (i.e., in order to take note of what the other person can see, the participant needs to see something themself). Scholars thus manipulated alignment with an own representation in order to assess other-related representation abilities (in a praiseworthy effort not to make conclusions about this from socially aligning conditions only). These fields could not start out with designs that assess social representation directly, like the other domains did. Yet, this difference in methodological limitations may not reflect actual phenomenological differences necessarily: For instance, in everyday social interactions, one will *think* but also *move* differently from others at various points in time. This does not mean that other-related action representation requires this movement difference to occur (Iacoboni & Dapretto, 2006; Rizzolatti & Craighero, 2004; Rizzolatti et al., 2001), nor that other-related mental representation should be impossible at those (perhaps few) moments where we do not hold a particular (conflicting) mental state ourself (e.g., Dziobek et al., 2006). Aligning the way we understand social conflict tasks under a relational framework, we will argue, may help to better understand how different social-cognitive domains may tie together in a shared mechanism. We summarised the characteristics of the different domains in Table 6.

Methodological equivalence to the Action Perception Domain: The empathy and Touch Perception Domain

In many respects, the domain of empathy has followed a history that is comparable to the action perception domain. Here, researchers started to investigate empathy by looking at how the human brain responds to directly observing someone in pain, while being in a neutral

state him- or herself, and typically contrasted this to control condition where neither the participant nor the observed agent experiences pain (for a recent review, see Fallon, Roberts, & Stancak, 2018). This design follows the action observation paradigm that is used in the mirror neuron literature focussing on action representations. Neuroimaging studies using this kind of design have suggested that the bilateral anterior insula and the anterior mid-cingulate are involved in the representation of others' pain (for reviews, see Fallon, Roberts, & Stancak, 2018; Lamm, Decety, & Singer, 2011). Given the involvement of these areas in the experience of one's *own* pain as well, these brain areas are thought to be part of the mirror neuron system.

In addition, researchers have only recently started to look into empathy *conflict* paradigms, in which the emotional state of the participant is experimentally manipulated as well as that of another person they should empathise with so that they are either congruent or incongruent (Hoffmann, Singer, & Steinbeis, 2015; Silani, Lamm, Ruff, & Singer, 2013; Steinbeis, Bernhardt, & Singer, 2015; von Mohr, Finotti, Ambroziak, & Tsakiris, 2019). Scholars have, for instance, presented participants with tactile stimuli that could leave them either disgusted or pleased (like a rotten apple versus a feather, respectively), while showing them another person who simultaneously experienced a similar tactile sensation that could elicit feelings of either disgust or pleasure (Silani et al., 2013). It is found that if the participant experiences an emotion opposite to that of the other following a tactile stimulation, ratings of the strength of the respective valence of the stimuli are drawn towards the emotion experienced by oneself, as compared to when they are both the same (Silani et al., 2013). Such an egocentric bias is seen as a reflection of the brain processing the social conflict that exists in the incongruent empathy condition. Noteworthy, it was reported that in such paradigms mostly the right supramarginal gyrus (rSMA), just anterior to the TPJ, is involved in the

processing of incongruent versus congruent empathy trials (Lamm, Bukowski, & Silani, 2016; Silani et al., 2013; Steinbeis et al., 2015). Further research could follow up on this nascent line of research.

Hence, not experiencing the methodological limitation mentalising researchers face, the empathy domain has typically followed a *representational* interpretation for pain *observation* studies, and a *relational* interpretation for empathy *conflict* studies, comparably to the interpretations that exist in the action perception domain). Similarly, the touch perception domain has achieved to differentiated knowledge about touch representation and touch conflict monitoring respectively from designs that focus on observing another's touch experience (e.g., Gazzola & Keysers, 2009; Keysers et al., 2004, 2010), and those implementing conflict between the location of own and the other's observed touch (e.g., Deschrijver, Wiersema, & Brass, 2016, 2017).

Methodological Equivalence to the Theory of Mind Domain: Perspective Taking, Moral Dilemmas, Lie Detection, Irony, Sarcasm and Humour

Perspective taking has long been identified as an important social-cognitive skill next to mental state representation. This domain was initialised by the development of a social congruency paradigm: In the well-known Director task (Keysar, Barr, Balin, & Brauner, 2000), participants listen to auditory instructions from another person (the 'director'), who asks them to move particular objects in a specific direction. In some cases, the relevant objects are only visible to the participant and hence to be ignored (i.e., competitor objects) as they cannot be the target object that the director is referring to. The task usually compares looking times and erroneous behaviour in the condition with a 'competitor' object (where the

perspective of the director on the target object is incongruent with that of the participant) versus that in which there is no 'competitor' object (where the perspective of the director on the target object is congruent with that of the participant). Just like in the Theory of Mind domain, dependent measures in these perspective conflict tasks are often (though not always) interpreted in terms of perspective *representation* (i.e., the ability to infer or 'take' the perspective of the director). Similarly, a lack of effect in these measures is sometimes thought of as signifying that the participant did not *represent* of the other's perspective. The design used in the perspective taking domain is however methodologically comparable to that of the Theory of Mind domain: Acting upon what the other can see (Keysar et al., 2000; Qureshi, Apperly, & Samson, 2010; Zwickel et al., 2011) may in part depend on conflict monitoring mechanisms after representation, rather than the latter per se. If own- and other-related perspective representations are present in both conditions, a neuroimaging contrast of an incongruent versus congruent perspective taking trials could isolate activity related to the *change in alignment* between the two perspectives, not to the *representation of* the other's perspective per se (i.e., the perspective *taking*; e.g., Schurz et al., 2013). Correspondingly, incongruent (versus congruent) perspective taking trials have consistently revealed activity in the TPJ (Schurz et al., 2013) and neuromodulation of the area interferes with performance in the task (Nobusako et al., 2017). This suggests that also here, the TPJ may act as a common social conflict monitoring mechanism taking place after representation has occurred rather than as a representational one, while the neural locus of perspective representation as such is less clear. Even the term 'perspective taking', which primarily refers to the *representation* of a perspective, may perhaps be better phrased in terms of 'perspective conflict monitoring', as paradigms that methodologically isolate perspective representation *per se* require more development.

Moral decision making is often discussed in terms of representational mechanisms, but investigated using tasks that can be understood in terms of social conflict processes instead. In a seminal study on moral judgements (Young, Cushman, Hauser, & Saxe, 2007), participants were confronted with a situation where a protagonist puts powder in a friend's coffee believing that it is either sugar or toxic (i.e., neutral or negative belief state), after which the friend is either fine or dies (i.e., neutral or negative outcome). The authors showed that the condition of attempted harm (i.e., the protagonist believes the powder to be toxic but the friend is fine nevertheless) elicited more activity than the other three conditions in the TPJ (see figure 4 of the paper's supplementary materials). Similarly, participants with a higher TPJ-response to accidental harms (neutral belief – negative outcome) are more forgiving and attribute less blame for accidents, compared to participants with a lower TPJ-response (Young & Saxe, 2009). Disruption by neuromodulation of the right TPJ relative to control site in the aforementioned paradigm, in addition, leads participants to specifically judge attempted harms (negative belief – neutral outcome) as less morally forbidden (Young, Albert, Hauser, Pascual-leone, & Saxe, 2010). The TPJ activation was thus tied specifically to conditions that implemented mental conflict, such as attempted harms (negative belief – neutral outcome) and accidents (neutral belief – negative outcome). The authors stressed the necessity to *represent* the other person's belief in these false belief situations, to decide how morally acceptable the situations were. However, this design is crucially different from a typical false belief task: In *each* of the conditions the other's mental state is *readily stated* ('Grace thinks that the powder is toxic' or 'Grace thinks that the powder is sugar'). If the TPJ would be tied to other-related *representation* per se, activity in the area could be expected across all conditions. Instead, TPJ involvement only being tied to attempted harms and accidents in both neuroimaging and neuromodulation studies suggest that those participants who process the *mismatch* between their knowledge of the character's belief (e.g., that the powder was sugar) and the outcome of

their action in reality (e.g., the friend dying nevertheless) are more willing to grant the character forgiveness than those who do not process this mismatch. In sum, we suggest that these data emphasise that detecting the extent to which our understanding of a person's motivations *conflicts* with our perception of its consequences in the world contributes to our sense of morality, rather than mental representation per se (Young et al., 2010, 2007; Young & Saxe, 2009).

Representation of others' mental states has in the past been suggested to play an important role in understanding complex language such as irony, lies (Sowden, Wright, Banissy, & Catmur, 2015) and humour (Samson et al., 2008; Samson, Hempelmann, Huber, & Zysset, 2009). In this context, the role of the TPJ has also been suggested to be in mental representation rather than in mental conflict monitoring. Yet, research on humour has a long history of incongruity theories (Ritchie, 2009), which assert that one needs to at least partially resolve an incongruity in order to understand the punch line of a joke. For instance, incongruity can arise from a clash between a certain set-up of the joke and the punchline (e.g., O'Riley was on trial for armed robbery. The jury came out and announced, "Not guilty". "Wonderful", said O'Riley, "does that mean I can keep the money ?"; Ritchie, 2009). Also here, one could assert that a difference is present between the world as understood by the other (i.e., the jury thinking that O'Riley is not guilty) and the actual state of the world outlined in the punchline (i.e., him actually being guilty). The task to understand this type of humour is thus one of social conflict monitoring (in TPJ) rather than of representation or mental state inference *per se* (note that also here, the jury's mental state is readily given rather than inferred). Not surprisingly from this perspective, the processing of incongruity-resolution cartoons, in contrast to nonsense cartoons that did not involve social conflict, leads to more activation in areas around the TPJ bilaterally (Samson et al., 2008, 2009). Activation in TPJ

is similarly linked to recognition of ironic communicative intentions (Bosco, Parola, Valentini, & Morese, 2017), where language is used to make a remark that mismatches the actual state of the world as understood by the listener (e.g., remarking that it is a beautiful day after rain and thunder suddenly appear), and for the detection of lies when the verbal information given by the speaker is inconsistent with one's own knowledge of the truth (Sowden, Wright, et al., 2015). The functional relationship between the TPJ and the understanding of irony, lies, and certain types of humour may consist of an appreciation of the interplay between different types of social information verbally given by others and the world as understood by oneself – rather than reflecting the attribution of mental states to others per se. In sum, to understand that someone is joking, lying, ironic or sarcastic, one strictly speaking doesn't need to infer what the other person is thinking: One can just listen and detect whether the verbal information provided by the other (i.e., the social information) mismatches one's own knowledge of the world. If one fails to detect such mismatch, one will not recognise a joke, lie, or ironic/sarcastic remark for what it is.

Overall, in this Section, we reviewed that in several social-cognitive domains that have had no history of implementing social representation designs before social conflict designs, effects obtained in social conflict designs have typically been interpreted in terms of other-related representation. By reframing results in terms of social conflict monitoring rather than other-related mental representation when appropriate, the field would achieve consistency over different social-cognitive domains in how social conflict designs are typically interpreted, while our understanding of the processes involved significantly changes. A focus of research going forward may be to use a socially congruent condition as a baseline for isolating activity related to social conflict, across different domains of social cognition.

Tying the Different Domains of Social Cognition Together

In a fairly recent line of research, scholars have started to functionally relate the different fields of human social cognition to one another, suggesting that a mechanism within the TPJ may be common to them all (Brass et al., 2009; Hogeveen et al., 2014; Santiesteban, White, et al., 2012; Santiesteban et al., 2015; Santiesteban, Banissy, et al., 2012; Sowden & Catmur, 2015; Sowden & Shah, 2014; Sowden, Wright, et al., 2015; Spengler et al., 2010, 2009). Yet, this tying together of different research domains has not always been with success so far, perhaps in part because of the hybrid theoretical framework implicit to the domain that places the role of conflict monitoring both prior and subsequent to mental representation.

In one study (Santiesteban, White, et al., 2012), for instance, participants were trained to either imitate (i.e., represent other-related actions) or inhibit imitation (i.e., to monitor action conflict). It was hypothesised that training action conflict monitoring – but not training action representation – should positively affect people's performance in a perspective taking task (the Director task) and a mentalising task not based on the Sally-Anne design (the triangles task). Such a hypothesis reflects an idea typical to a hybrid framework that lower-level action conflict monitoring abilities (reflected in the action conflict monitoring paradigm) support the supposedly more high-level ability of *representing* others in mentalising and perspective taking tasks (Sowden & Shah, 2014). The group trained to monitor action conflict showed improved performance in measures of the Director task, yet no effect was found for belief verbalisation measures in the triangle mentalising task (Santiesteban et al., 2012). Imitation training (i.e., training action representation) did not have any effects on these measures. In accordance with the motivating assumption outlined above, the authors concluded that training action conflict monitoring enhances 'the ability to adopt the perspective of others' (i.e., *representing* the other's perspective after dealing with perspective

conflict), while suggesting that ceiling effects accounted for the null result in the verbalisation measures of the mentalising task (i.e., in *representing* the other's mental state). However, as noted above, the dependent measures of the Director task (Keysar et al., 2000) involve a comparison of socially incongruent versus congruent eyetracking data (eyetracking of 'competitor' versus 'non-competitor' objects), which we have argued reflect perspective *conflict* monitoring. Arguably, these results instead indicate that action conflict monitoring training enhances the participants' *processing of perspective conflict*. In addition, the belief verbalisation task can be seen as implementing a dependent measure focussed on expressing the other-related mental representation in task that may involve latent mental conflict, meaning that it may capture other-related representation more strongly than the other paradigms. If the imitation inhibition training paradigm influences relational mechanisms in TPJ, one would not necessarily expect an effect on mental representation per se, as there may be different representational systems for the representation of others' actions versus mental states (i.e., respectively the somatosensory and motor mirror neuron system versus so far unidentified areas for mental states). Similarly, one should not expect an effect of training action *representation* on a mental representation measure.

A number of studies have targeted the TPJ via neuromodulation techniques in order to investigate functional relationships between different social-cognitive domains. Here, it was shown that neuromodulation of the activity in the TPJ influences action conflict monitoring (with transcranial magnetic stimulation, tDCS, Hogeveen et al., 2014; Nobusako et al., 2017; Santiesteban et al., 2015; as well as with transcranial magnetic stimulation, TMS, Sowden & Shah, 2014), 'competitor' versus 'non-competitor' condition comparisons in perspective taking tasks (i.e., perspective conflict monitoring in the framework of the current paper, with tDCS, Nobusako et al., 2017; as well as with TMS, Santiesteban et al., 2015; Santiesteban,

Banissy, et al., 2012), and performance in lie detection (i.e., monitoring socially inconsistent situations, with tDCS, Sowden, Wright, et al., 2015). The neuromodulation did not affect action imitation (i.e., action representation, with tDCS, Hogeveen et al., 2014) nor mental judgements about oneself or others (i.e., representation of own and other-related mental states, with TMS, Santiesteban et al., 2015; Santiesteban, Banissy, et al., 2012). This pattern of findings is consistent with the notion that TPJ is involved in conflicting monitoring processes more so than mental representation per se. Specifically, modulation of the TPJ appears to affect social conflict measures (of action-, perspective-, or mental conflict) but not representational measures (of actions or mental states per se). In sum, we think a careful application of social conflict interpretations for (latent) mental conflict measures, and social representation interpretations for representation-only measures, more parsimoniously explains why certain social-cognitive findings tie together, and why others do not.

Overall, the neural substrates of social cognition have in the most influential meta-analytic and theoretical papers been framed in terms of two distinct large-scale *representational* networks: The mirror neuron system, reflecting the *representation* of others' actions in premotor and somatosensory areas amongst others (Rizzolatti & Craighero, 2004), and the mentalising network (Van Overwalle, 2009; Schaafsma, Pfaff, Spunt, & Adolphs, 2016; Schilbach et al., 2013), reflecting *representation* of others' mental states in TPJ-areas (Apperly et al., 2004; Lombardo et al., 2007, 2011; Overwalle, 2009; Samson et al., 2004; Saxe & Kanwisher, 2003; Saxe & Powell, 2006)). However, the neural mechanisms of social cognition can be re-considered in light of the distinction between representational and relational social cognition. Specifically, the neural substrates of social cognition may rather consist of *one* relational hub (i.e., the TPJ-area), which subserves *a multitude* of distinct large-scale representational networks (i.e., of mental representation, perspective representation,

etc.). This view follows earlier scholars that pointed out that different social-cognitive domains may share a common mechanism in TPJ (Brass et al., 2009; Hogeveen et al., 2014; Santiesteban, White, et al., 2012; Santiesteban et al., 2015; Santiesteban, Banissy, et al., 2012; Sowden & Catmur, 2015; Sowden & Shah, 2014; Sowden, Wright, et al., 2015; Spengler et al., 2010, 2009), but also contrasts with a notion where this is seen as a lower-level self-other distinction mechanism supporting a higher-level *representation* of mental states and perspectives located in the same area (Van Overwalle & Baetens, 2009; Schurz et al., 2013; Sowden & Shah, 2014). Note that the domains of moral reasoning, humour, irony, sarcasm, and lie detection all use designs where the other's mental state is readily given (rather than inferred), and still consistently report TPJ involvement. This suggests that a common mechanism may compare any understanding of (verbally given or inferred) social information with one's own information. From this view, it follows that social conflict monitoring is the 'higher-level' skill in human social cognition, rather than social representation per se. If monitoring conflict, the TPJ area probably receives input from many different sources. For example, when monitoring action conflict, it may receive information from the motor cortex, whereas when monitoring mental conflict, it may receive information from (currently unidentified) areas involved in mental representation different from those involved in action representation. Those two sources of information are presumably connected to the TPJ through different neural pathways. As a consequence, one can imagine that the conflict resulting from those very different sources of information would be decodable within the TPJ. In the same regard, and on a more specific level, we think that properties of beliefs that are potentially represented in different areas or encoded differentially within certain areas of the brain (such as the perceptual source of information or strength of evidence; Koster-hale et al., 2017; Koster-Hale, Bedny, & Saxe, 2014) might yield differentially decodable conflict within TPJ after following distinct neural pathways that lead to this area or after providing

differentially decodable input. What this shows, is that the neural activations related to conflict monitoring may not be uniform across contexts, but are dependent on the inputs to the monitoring mechanism.

Social Conflict Monitoring in Everyday Life

With over 7000 citations for the most seminal of Sally-Anne studies (Baron-Cohen et al., 1985), the understanding of social conflict tasks, and of social cognition altogether, in terms of inferring mental representations, Theory of Mind or 'mindreading' (Apperly, 2010; Carruthers, 2013, 2016, 2017; Dziobek et al., 2006; Thompson, 2014) has spread well beyond the mentalising domain alone. In autism, an idea that social difficulties are tied to 'mindblindness' or an inability to infer and afterwards represent others' mental states has become almost factual in popular and scientific understanding. We have instead argued that the ability to neurally determine the correspondence of another's social information with our own may provide us with the most essential information for social understanding. No two people ever see or experience the same events in exactly the same light: Human communication and understanding may not depend so much on inferring someone else's mental state as such, but rather on getting an impression of how well our present information about them aligns with that of ourselves. An individual that does not fully process the difference with others' line of thought, in contrast, may experience social difficulties.

The enormous attention for mindreading in mentalising research (e.g., Apperly, 2010)

has followed from early theoretical and empirical arguments (Baron-Cohen et al., 1985; Dennett, 1978; Premack & Woodruff, 1978) that the human mind is able to get to the content of what another individual is *thinking* by taking into account what this person has *perceived*. In certain social circumstances, one may indeed infer and use the content of another individual's mental state in a social interaction: When you observe your company in a bar looking at an empty glass, for instance, it is reasonable to assume that this person is thinking about getting the glass refilled. In many other instances, however, the exact belief content of your partner will likely be difficult to *just* infer. In fact, research has shown that people are not much better than a coin toss when trying to discriminate others' lies from truths if they do not hold the truth themselves (Bond & DePaulo, 2006), suggesting that actual 'mindreading' is a near impossible in some circumstances. Luckily, in many situations, access to another person's truthful appreciation of the world is readily available via their speech, bypassing the need for inference *per se*. The ability to *use different types of social information* about the individual (actions, touch, pain, perspectives, verbally stated or inferred mental states) to assess whether an individual is on the same wavelength, may here be more advantageous than an ability to infer mental states indirectly. In this sense, sensory cues in social behaviour (e.g., in facial expressions) may also signal when another person is not on the same page, without revealing *what* it is exactly that this person is thinking. A person's misunderstanding of our line of reasoning may signal a need to rephrase our own thoughts or make a move towards the other's – regardless of knowing the exact content of the misunderstanding. Social cognition may thus depend more on how the mental state or other experience of the other person strikes us *with regard to our own* (i.e., as a relational process) rather than how we *infer* them (i.e., as a representational process).

In individuals on the autism spectrum, social difficulties can range from an absence of

reciprocity within conversations, to seemingly inconsiderate or 'rude' behaviours towards others and difficulties in understanding social relationships such as friendship and love (American Psychiatric Association, 2013). Communicative differences are sometimes characterised by a lesser understanding of concepts such as irony, sarcasm and lies. It is not difficult to see how a lesser active social conflict monitoring could contribute to the social difficulties characteristic of autism. If less able to track the divergence of another's mental state with one's own after representing it, one may become less sensitive to the rhythm of social conversation: One may not notice that the context actually requires one to focus on the other person's understanding of things, resulting in one persevering in one's own line of thought. For instance, when a person tries to steer away the topic of conversation to their own interest, individuals on the autism spectrum may not process this social conflict as a cue to inhibit themselves in talking about their own interests. The reverse may occur as well: Understanding another's mental state may interfere with a person on the autism spectrum's own knowledge and beliefs, leading them to experience inhibition to verbalise these even if that were most appropriate for the context. Such an underlying mechanism clearly differs from individuals on the spectrum being blind to others' mental states altogether, and these exact differences in social reciprocity and back-and-forth conversation are amongst the core features of autism (American Psychiatric Association, 2013). Even our ability to befriend someone, or fall in love with someone, may ultimately rely on detecting mental (mis)alignment: Repeated interaction with a person can allow us to more gradually develop an impression of whether this person is generally *like-minded* or not (Parkinson, Kleinbaum, & Wheatley, 2018), leading us to understand the nature of our relationship. Over longer timescales of mental conflict monitoring difficulties, in contrast, one may naturally end up feeling socially alienated rather than truly connected. In this sense, lesser mental conflict monitoring may be much subtler a social difficulty than pure 'mindblindness' (Baron-Cohen

et al., 1985; Lombardo & Baron-Cohen, 2011; Lombardo et al., 2011; Senju et al., 2009).

This framework may also shed light on clinical challenges of other populations: A multitude of patient groups other than autism show decreased TPJ activity or what may be other indications of mental conflict processing difficulties in social cognition tasks (e.g., in eyetracking measures), including those with a diagnosis of schizophrenia (Das, Lagopoulos, Coulston, Henderson, & Malhi, 2012; Kronbichler, Tschernegg, Martin, Schurz, & Kronbichler, 2017; Lee, Horan, Wynn, & Green, 2016; van der Weiden, Prikken, & van Haren, 2015), Parkinson's disease (Emre Bora, Walterfang, & Velakoulis, 2015), bipolar disorder (Bora, Bartholomeusz, & Pantelis, 2018; Bora & Pantelis, 2016), depression (Lee, Harkness, Sabbagh, & Jacobson, 2005), specific language impairment (Nilsson & de López, 2016) and eating disorders (Bora, 2016; Cazzato, Mian, Serino, Mele, & Urgesi, 2015). Our theoretical framework suggests that, at least for some social difficulties, an atypical mental conflict mechanism (but not necessarily lacking other-related mental representation or 'mindblindness') may lie at their heart. Though one always needs to be careful with deriving clinical instructions from basic research questions, our framework hints towards the importance of developing a clinical practice that treats mental conflict processing difficulties as relevant to everyday social issues.

Conclusions

Altogether, we have emphasised the role of *relational* aspects of social cognition, namely, for instance how we monitor the correspondence between our own and other people's mental states during social interactions, after our brain takes note of them. This is largely based on a relational reinterpretation of the past 40 years of false belief data in terms of mental

conflict monitoring rather than inference of another's mental representation. This reframing is important to how we understand experimental design in the study of social cognition, and the specific mechanisms that are revealed by neuroimaging data in this domain and which may underlie differences in behavioural performance and neural responses in clinical cohorts and across developmental trajectories. Based on this, we highlighted three key advances for the social-cognitive domain: We have argued specifically for the role of TPJ as a neural mechanism that potentially monitors mental conflict specifically, rather than enacting the representation of other people's mental states *per se*. Further, we have argued that individuals on the autism spectrum may not experience complete 'mindblindness', while atypical effects in false belief data of young children and of non-human primates may not indicate lacking representational abilities either. Reframing the core neural mechanisms of mentalising in this way has consequences for the larger social-cognitive domain, including in the fields of perspective taking, moral judgements, lie detection, humour, irony and sarcasm. A greater consistency in the use of relational interpretations for social conflict data holds promise for furthering our understanding of how different social-cognitive domains are tied together. Moreover, broader fields like legal topics (Kliemann, Young, Scholz, & Saxe, 2008), animal cognition (Call & Tomasello, 2008; Premack & Woodruff, 1978), cultural studies (Callaghan et al., 2005) and economics (Robalino & Robson, 2012) have adopted the traditional emphasis on mental state inference as core to how we understand others. In the long run, if the alternative framework that we propose holds, these disciplines may also need to accommodate a central role of relational mentalising in human social cognition. If mindreading per se is less instrumental for navigating the social world than previously thought, relational mentalising opens up potential new avenues for psychological science and beyond.

Tables

Basic assumptions	Representational framework	Relational framework
Core cognitive mechanism?	Inferring the other-related representation *e.g., Theory of Mind, mental state inference, 'mindreading' or perspective taking*	Social conflict monitoring *e.g., mental conflict monitoring*
Timing of social conflict monitoring?	Before other-related representation *e.g., perceptual conflict detection required for mental representation or for perspective taking*	After other-related representation *e.g., conflict monitoring needed after other's action is represented next to own action representation in mirror neuron areas*
Neural representation of socially congruent other?	Not necessarily *e.g., another's belief or perspective may not be represented when it matches with our own*	Yes *e.g., another's congruent action, perspective or belief is represented*
Dependent measures primarily reflect?	Other-related representation abilities	Social conflict monitoring abilities after representation (some focus more strongly on the expression of the own or other-related representation)
Aim of control condition?	Isolating other-related representation from non-mentalistic conflict processes *e.g., via physical conflict condition*	Isolating social conflict from preceding representational processes *e.g., via congruent action or true belief condition*
Likely cause of atypical effects in a socially incongruent situation?	Lacking other-related representation *e.g., atypical false belief or perspective taking result signifies a lack of other-related representation*	Inactive conflict monitoring mechanism (can follow from lacking other-related mental representation, but not necessarily) *e.g., atypical false belief result signifies an inactive mental conflict monitoring mechanism, which may have followed from a lacking other-related representation, but not necessarily*

Table 1. Basic assumptions for a representational versus relational framework in the context of a social congruency paradigm. Note. Examples are provided in italics.

	Population with lacking other-related representation+
All dependent measures reflect other-related mental representation*	No effect of other's mental state

Table 2. Predictions for performance in a false belief condition under the assumptions of a representational framework (i.e., in terms of them primarily reflecting mental representation).

*Dependent measures can be considered implicit (e.g., looking times) or explicit (e.g., belief verbalisations). It has been proposed that difficulties with other-related representation may be compensated for more easily in the latter compared to the former.
+Lacking representations may follow from ineffective detection of perceptual conflict.

	Population with lacking other-related representation	Population with ineffective mental conflict monitoring
Dependent measure focussed on expressing other's mental states (e.g., verbalising other's belief)	No effect of other's mental state	Less influence of other's mental state
Dependent measure focussed on expressing participant's own mental states (e.g., verbalising own belief or measuring expectations about a target's location with RTs)	No effect of other's mental state	More influence of other's mental state
Dependent measure without any inherent focus (e.g., neuroimaging)	No effect of other's mental state	No processing of mental conflict

Table 3. Predictions for performance in a false belief condition compared to a true belief condition under the assumptions of a relational interpretation for social congruency data (i.e., primarily reflecting social conflict).

Dependent measures used in a false belief condition reflect social conflict monitoring rather than mental representation per se. Lacking mental representations should lead to a consistently lacking effect of the other's mental state in data of all dependent measures. With an inactive conflict monitoring system, the influence of the other's mental state may vary, however, depending on the dependent measure used. We indicated in grey the situations where an inactive common conflict monitoring system could potentially lead to inconsistent results.

	Isolated process(es) under representational framework	**Isolated process(es) under relational framework**
True belief condition	Other-related representation coinciding with perceptual conflict monitoring	Mental conflict monitoring
Physical conflict condition ***e.g., false photograph condition***	Other-related representation	Other-related representation, own representation and mental conflict monitoring
Other conditions	Other-related representation coinciding with perceptual conflict monitoring	Other-related representation, own representation, mental conflict monitoring and (perhaps) perceptual conflict monitoring

Table 4. Control conditions most commonly used for the false belief condition in neuroimaging studies, and what they are thought to isolate under a representational versus relational interpretation framework.

We indicated in grey the control condition preferred under each respective framework.

Reference	Stimuli type	Explicit / Implicit	Contrast	TPJ activity	Other areas	Key finding reported
(Abraham, Rakoczy, Werning, von Cramon, & Schubotz, 2010); n = 22	Vignettes	Explicit	FB > TB	No	(ventral) mPFC, posterior cingulate, retrosplenial cortex, hippocampus, amygdala, insula, superior temporal gyrus, putamen, premotor cortex, precentral gyrus, inferior frontal gyrus	Incongruent intentional states (false beliefs/unfulfilled desires) versus congruent (true beliefs/fulfilled desires) involves medial wall
(Aichhorn et al., 2008); n = 21	Vignettes	Explicit	FB > TB	Yes	Anterior medial temporal gyrus, temporal pole	rTPJ is especially sensitive to processing belief information
(Bardi, Desmet, Nijhof, Wiersema, & Brass, 2017); n = 22	Movies	Implicit & explicit	FB > TB	Yes	Angulary gyrus, fusiform gyrus/collateral gyrus	Neural mechanisms for explicit and implicit Theory of mind overlap in TPJ and mPFC
(Boccadoro et al., 2019); n = 68	Movies	Implicit	FB > TB	Yes	Middle temporal gyrus, supramarginal gyrus, lingual gyrus, inferior frontal gyrus, inferior parietal lobule, middle frontal gyrus, medial orbitofrontal cortex, postcentral gyrus, thalamus/caudate nucleus, insula, vermis	Multi-study of implicit Theory of Mind reveals cluster of activation in posterior parietal cortex spanning the TPJ, but no mPFC activation
(Cracco et al., 2020); n = 31	Movies	Implicit	FB > TB	Yes	Middle temporal gyrus, temporal pole, dorsomedial prefrontal cortex, precuneus	Women with interpersonal trauma show less TPJ activation than controls
(Döhnel et al., 2012): n = 18	Cartoons	Explicit	FB > TB	No	Posterior mPFC/dorsal anterior cingulate cortex, dorsolateral prefrontal cortex, premotor cortex, dorsolateral prefrontal cortex, inferior frontal cortex, middle temporal gyrus, inferior parietal cortex/ superior parietal cortex, precuneus, thalamus	Common TPJ activity for true and false belief processing
			TB > FB	No	Anterior mPFC, ventromedial prefrontal cortex, inferior parietal cortex	
(Döhnel et al., 2017); n = 22	Cartoons	Explicit	FB > TB behaviour attribution	Yes	Frontopolar cortex, mPFC, dorsolateral prefrontal cortex, premotor cortex, inferior frontal cortex, middle temporal gyrus	Emotion and behaviour belief reasoning elicit both TPJ and MPFC, differential anterior dorsolateral prefrontal cortex for emotion attribution
(Kovács et al., 2014); n = 15	Movies	Implicit	FB > TB pos. content	Yes	mPFC	Implicit false beliefs also elicit TPJ and mPFC activity
(Nijhof et al., 2018); n = 21	Movies	Implicit & explicit	FB > TB	Yes	Angular gyrus, lingual gyrus, dorsolateral prefrontal cortex	Lower TPJ activity for false versus true beliefs in adults with autism during implicit and explicit task
(Özdem et al., 2017); n = 20	Movies	Explicit	FB > TB	Yes	Precuneus	Overlap in brain activity for belief task process (spatial versus verbal belief) and for the same spatial input modality (spatial belief versus spatial reorientation)

Reference	Stimuli type	Explicit / Implicit	Contrast	TPJ activity	Other areas	Key finding reported
(Özdem et al., 2019); n = 25	Movies	Explicit	Similar FB > Mixed TB > Mixed FB > Similar TB	Yes	(Posterior) medial frontal cortex, inferior frontal gyrus	An increasing number of agents holding a false belief increases activation in the TPJwhen participants have to report the belief of the agent.
(Rothmayr et al., 2011b); n = 12	Cartoons	Explicit	FB > TB	Yes	Middle frontal gyrus, precentral gyrus, medial frontal gyrus, middle temporal gyrus, thalamus, precuneus, superior frontal gyrus	Overlap for belief reasoning and inhibitory control in mPFC and TPJ
(Schneider, Slaughter, et al., 2014); n = 16	Movies	Implicit	FB > TB	No	Anterior superior temporal sulcus, precuneus	No difference in TPJ for implicit false versus true belief
(Sommer et al., 2007); n = 16	Cartoons	Explicit	FB > TB	Yes	Medial dorsal anterior cingulate cortex, middle frontal gyrus, dorsolateral prefrontal cortex, lateral rostral prefrontal cortex, middle temporal gyrus, inferior parietal gyrus, precuneus	Dorsal anterior cingulate cortex, TPJ and lateral prefrontal cortex involved in false versus true belief
			TB > FB	No	Medial rostral prefrontal cortex	
(Sommer et al., 2018); n = 15	Movies	Explicit	FB > TB question	No	Inferior frontal gyrus, inferior temporal gyrus	Increased TPJ activity in question phase for adults with autism
			TB > FB question	Yes	mPFC, posterior frontal gyrus, precuneus	
(Wysocka et al., 2020); n = 13	Movies	Explicit	FB > TB	No	Middle temporal gyrus	The task involves structures in the Theory of Mind network

Table 5. Studies that used a true belief condition as a control for the false belief condition.
We conducted a systematic review in the database Web of Knowledge. We performed a keyword search with the keywords "theory of mind', "mindreading", "false belief", "mentalising" or "mentalizing" and one of the keywords "neuroimaging" "functional magnetic resonance", "fMRI", "positron emission tomography", or "PET". We focussed on studies that have used a typical false belief manipulation, where an agent has a belief that conflicts with the participant's and with reality. Studies were only selected if they experimentally contrasted a socially incongruent (false) belief condition against a socially congruent (true) belief condition. We did not select studies which focused on 'second order' false beliefs ('x thinks that y thinks that…'). The search was limited to studies published until May 2020. If a study included a clinical group next to a neurotypical one, we included results of the contrast in the neurotypical control group. If a study reported more than one contrast involving a false and true belief condition, we selected the one best corresponding to contrasts reported in the other studies. Mind that the specific contrast selected may not have been the focus of the respective study, such that the results from the false belief versus true belief contrast we focussed on usually differ from the 'key finding' as noted in the study, which we also reported in the table. FB = false belief, TB = true belief, mPFC = medial prefrontal cortex; TPJ = temporoparietal junction. Mind that Boccadoro et al., (2019) report in part on datasets also discussed by Bardi et al. (2017) and Nijhof et al. (2018). Note that the studies listed in the table differ in whether they employ a fully balanced design or not.

	Conceptual focus typical for the domain	Conceptual focus under a relational framework
Action conflict domain	Action conflict	Action conflict
Touch conflict domain	Touch conflict	Touch conflict
Empathy conflict domain	Pain conflict	Pain conflict
Theory of Mind domain	Other-related mental representation	Mental conflict
Perspective taking domain	Often other-related perspective representation	Perspective conflict
Morality domain	Other-related mental representation	Conflict between other's intention/belief statement and the self-perceived outcome of other's action in the world
Lie detection domain	Other-related mental representation	Conflict between other's verbal statement and own knowledge of truth
Irony/Sarcasm domain	Other-related mental representation	Conflict between other's verbal statement and own knowledge of actual state of the world
Humour domain	Other-related mental representation	Conflict between other's statement and own knowledge of actual state of the world in the punchline

Table 6. Social-cognitive domains that have a social congruency paradigm as their main methodological design, their typical focus for interpreting a socially incongruent condition, and the conceptual focus of interpretation under a relational framework.

We indicated in grey the domains where the typical conceptual focus for data interpretation aligns with the methodological focus of the paradigm already, that is, in social conflict.

References

Abraham, A., Rakoczy, H., Werning, M., von Cramon, D. Y., & Schubotz, R. I. (2010). Matching mind to world and vice versa: Functional dissociations between belief and desire mental state processing. *Social Neuroscience*, *5*(1), 1–18. https://doi.org/10.1080/17470910903166853

Aichhorn, M., Perner, J., Weiss, B., Kronbichler, M., Staffen, W., & Ladurner, G. (2008). Temporo-parietal Junction Activity in Theory-of-Mind Tasks : Falseness , Beliefs , or Attention. *Journal of Cognitive Neuroscience*, *21*(6), 1179–1192.

Alcalá-López, D., Vogeley, K., Binkofski, F., & Bzdok, D. (2019). Building blocks of social cognition: Mirror, mentalize, share? *Cortex*, *118*, 4–18. https://doi.org/10.1016/j.cortex.2018.05.006

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (5th ed.)*. *Arlington, VA: American Psychiatric Publishing. Washington, DC: Author.*

Apperly, I. A. (2010). *Mindreaders: The Cognitive Basis of "Theory of Mind"* (1st ed.). Psychology Press.

Apperly, I. A., & Butterfill, S. A. (2009). Do Humans Have Two Systems to Track Beliefs and Belief-Like States ?, *116*(4), 953–970. https://doi.org/10.1037/a0016923

Apperly, I. A., Samson, D., Chiavarino, C., & Humphreys, G. W. (2004). Frontal and temporo-parietal lobe contributions to theory of mind: Neuropsychological evidence from a false-belief task with reduced language and executive demands. *Journal of Cognitive Neuroscience*, *16*(10), 1773–1784. https://doi.org/10.1162/0898929042947928

Baillargeon, R., Scott, R. M., & He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, *14*(3), 110–118. https://doi.org/10.1016/j.tics.2009.12.006

Banissy, M. J., & Ward, J. (2013). Mechanisms of self-other representations and vicarious experiences of touch in mirror-touch synesthesia. *Frontiers in Human Neuroscience*, *7*(APR 2013), 1–3. https://doi.org/10.3389/fnhum.2013.00112

Bardi, L., Desmet, C., Nijhof, A., Wiersema, J. R., & Brass, M. (2016). Brain activation for spontaneous and explicit false belief tasks overlaps: new fMRI evidence on belief processing and violation of expectation. *Social Cognitive and Affective Neuroscience*, *12*(3), 391–400.

Bardi, L., Desmet, C., Nijhof, A., Wiersema, J. R., & Brass, M. (2017). Brain activation for spontaneous and explicit false belief tasks overlaps: New fMRI evidence on belief processing and violation of expectation. *Social Cognitive and Affective Neuroscience*, *12*(3), 391–400. https://doi.org/10.1093/scan/nsw143

Bardi, L., Six, P., & Brass, M. (2018). Repetitive TMS of the temporo-parietal junction disrupts participant's expectations in a spontaneous Theory of Mind task, (February), 1775–1782. https://doi.org/10.1093/scan/nsx109

Baron-Cohen, S., Jolliffe, T., Mortimore, C., & Robertson, M. (1997). Another Advanced Test of Theory of Mind: Evidence from Very High Functioning Adults with Autism or Asperger Syndrome. *Journal of Child Psychology and Psychiatry*, *38*(7), 813–822. https://doi.org/10.1111/j.1469-7610.1997.tb01599.x

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition*, *21*, 37–46. https://doi.org/10.1016/0010-0277(85)90022-8

Bloom, P., & German, T. P. (2000). Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*, *77*(1), 25–31. https://doi.org/10.1016/S0010-

0277(00)00096-2
Boccadoro, S., Cracco, E., Hudson, A. R., Bardi, L., Nijhof, A. D., Wiersema, J. R., … Mueller, S. C. (2019). Defining the neural correlates of spontaneous theory of mind (ToM): An fMRI multi-study investigation. *NeuroImage*, *203*(September), 116193. https://doi.org/10.1016/j.neuroimage.2019.116193
Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, *10*(3), 214–234. https://doi.org/10.1207/s15327957pspr1003_2
Bora, E. (2016). Meta-Analysis of Theory of Mind in Anorexia Nervosa and Bulimia Nervosa : A Specific Impairment of Cognitive Perspective Taking in Anorexia Nervosa ?, (July), 739–749. https://doi.org/10.1002/eat.22572
Bora, E., Bartholomeusz, C., & Pantelis, C. (2018). Meta-analysis of Theory of Mind (ToM) impairment in bipolar disorder, (2016), 253–264. https://doi.org/10.1017/S0033291715001993
Bora, E., & Pantelis, C. (2016). Social cognition in schizophrenia in comparison to bipolar disorder : A meta-analysis. *Schizophrenia Research*, *175*(1–3), 72–78. https://doi.org/10.1016/j.schres.2016.04.018
Bora, E., Walterfang, M., & Velakoulis, D. (2015). Theory of mind in Parkinson ' s disease : A meta-analysis. *Behavioural Brain Research*, *292*, 515–520. https://doi.org/10.1016/j.bbr.2015.07.012
Bosco, F. M., Parola, A., Valentini, M. C., & Morese, R. (2017). Neural correlates underlying the comprehension of deceitful and ironic communicative intentions. *CORTEX*, *94*, 73–86. https://doi.org/10.1016/j.cortex.2017.06.010
Bowman, L. C. (2015). Children's belief- and desire-reasoning in the temporoparietal junction : evidence for specialization from functional near-infrared spectroscopy, *9*(October), 1–12. https://doi.org/10.3389/fnhum.2015.00560
Brass, M., Bekkering, H., & Prinz, W. (2001). Movement observation affects movement execution in a simple response task. *Acta Psychologica*, *106*, 3–22. https://doi.org/10.1016/S0001-6918(00)00024-X
Brass, M., Bekkering, H., Wohlschläger, A., & Prinz, W. (2000). Compatibility between observed and executed finger movements: comparing symbolic, spatial, and imitative cues. *Brain and Cognition*, *44*, 124–143. https://doi.org/10.1006/brcg.2000.1225
Brass, M., Derrfuss, J., & Von Cramon, D. Y. (2005). The inhibition of imitative and overlearned responses: A functional double dissociation. *Neuropsychologia*, *43*, 89–98. https://doi.org/10.1016/j.neuropsychologia.2004.06.018
Brass, M., Ruby, P., & Spengler, S. (2009). Inhibition of imitative behaviour and social cognition. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*, 2359–2367. https://doi.org/10.1098/rstb.2009.0066
Brass, M., Zysset, S., & von Cramon, D. Y. (2001). The inhibition of imitative response tendencies. *NeuroImage*, *14*, 1416–1423. https://doi.org/10.1006/nimg.2001.0944
Burnside, K., Wright, K., & Poulin-dubois, D. (2017). Social Motivation and Implicit Theory of Mind in Children With Autism Spectrum Disorder, (August), 1834–1844. https://doi.org/10.1002/aur.1836
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. https://doi.org/10.1016/j.cognition.2009.05.006
Butterfill, S. A., & Apperly, I. A. N. A. (2013). How to Construct a Minimal Theory of Mind, *28*(5), 606–637.
Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences*, *12*(5), 187–192.

https://doi.org/10.1016/j.tics.2008.02.010
Callaghan, T., Rochat, P., Lillard, A., Claux, M. L., Odden, H., Itakura, S., … Singh, S. (2005). Synchrony in the Onset of Mental State Reasoning. *Psychological Science*, *16*(5), 378–384.
Carlson, S. M. (2002). How Specific is the Relation between Executive Function and Theory of Mind? Contributions of Inhibitory Control and Working Memory. *Infant and Child Development*, *11*, 73–92. https://doi.org/10.1002/icd
Carlson, S. M. (2010). Developmentally Sensitive Measures of Executive Function in Preschool Children. *Developmental Neuropsychology*, *28*(October 2011), 37–41. https://doi.org/10.1207/s15326942dn2802
Carlson, S. M., Mandell, D. J., & Williams, L. (2004). Executive function and theory of mind: Stability and prediction from ages 2 to 3. *Developmental Psychology*, *40*(6), 1105–1122. https://doi.org/10.1037/0012-1649.40.6.1105
Carlson, S. M., & Moses, L. J. (2001). Individual Differences in Inhibitory Control and Children's Theory of Mind. *Child Development*, *72*(4), 1032–1053. https://doi.org/10.1111/1467-8624.00333
Carlson, S. M., Moses, L. J., & Claxton, L. J. (2004). Individual differences in executive functioning and theory of mind: An investigation of inhibitory control and planning ability. *Journal of Experimental Child Psychology*, *87*(4), 299–319. https://doi.org/10.1016/j.jecp.2004.01.002
Carruthers, P. (2013). Mindreading in infancy. *Mind and Language*, *28*(2), 141–172. https://doi.org/10.1111/mila.12014
Carruthers, P. (2016). Two Systems for Mindreading? *Review of Philosophy and Psychology*, *7*, 141–162. https://doi.org/10.1007/s13164-015-0259-y
Carruthers, P. (2017). Mindreading in adults : evaluating two-systems views. *Synthese*, *194*(3), 673–688. https://doi.org/10.1007/s11229-015-0792-3
Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement of mind : a functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage*, *12*(3), 314–325. https://doi.org/10.1006/nimg.2000.0612
Cazzato, V., Mian, E., Serino, A., Mele, S., & Urgesi, C. (2015). Distinct contributions of extrastriate body area and temporoparietal junction in perceiving one's own and others' body. *Cognitive, Affective & Behavioral Neuroscience*, *15*(1), 211–28. https://doi.org/10.3758/s13415-014-0312-9
Cook, J. L. (2014). Task-relevance dependent gradients in medial prefrontal and temporoparietal cortices suggest solutions to paradoxes concerning self/other control. *Neuroscience and Biobehavioral Reviews*, *42*, 298–302. https://doi.org/10.1016/j.neubiorev.2014.02.007
Cracco, E., Bardi, L., Desmet, C., Genschow, O., Rigoni, D., De Coster, L., … Brass, M. (2018). Automatic imitation: a meta-analysis. *Psychological Bulletin*.
Cracco, E., Bardi, L., Desmet, C., Genschow, O., Rigoni, D., De Coster, L., … Brass, M. (2018). Automatic Imitation: A Meta-Analysis. *Psychological Bulletin*. https://doi.org/10.1037/bul0000143
Cracco, E., Hudson, A. R., Van Hamme, C., Maeyens, L., Brass, M., & Mueller, S. C. (2020). Early interpersonal trauma reduces temporoparietal junction activity during spontaneous mentalising. *Social Cognitive and Affective Neuroscience*, *15*(1), 12–22. https://doi.org/10.1093/scan/nsaa015
Cross, K. a., Torrisi, S., Reynolds Losin, E. a., & Iacoboni, M. (2013). Controlling automatic imitative tendencies: Interactions between mirror neuron and cognitive control systems. *NeuroImage*, *83*, 493–504. https://doi.org/10.1016/j.neuroimage.2013.06.060

Das, P., Lagopoulos, J., Coulston, C. M., Henderson, A. F., & Malhi, G. S. (2012). Mentalizing impairment in schizophrenia : A functional MRI study. *Schizophrenia Research*, *134*(2–3), 158–164. https://doi.org/10.1016/j.schres.2011.08.019

de Guzman, M., Bird, G., Banissy, M. J., & Catmur, C. (2016). Self–other control processes in social cognition: From imitation to empathy. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1686). https://doi.org/10.1098/rstb.2015.0079

Decety, J., & Lamm, C. (2007). The Role of the Right Temporoparietal Junction in Social Interaction : How Low-Level Computational Processes Contribute to Meta-Cognition, *The Neuroscientist*, *13*(6), 580–593. https://doi.org/10.1177/1073858407304654

Dennett, D. C. (1978). Beliefs about beliefs [P&W, SR&B]. *Behavioral and Brain Sciences*, *1*(4), 568–570. https://doi.org/10.1017/S0140525X00076664

Deschrijver, E., Bardi, L., Wiersema, J. R., & Brass, M. (2015). Behavioral measures of implicit theory of mind in adults with high functioning autism. *Cognitive Neuroscience*, *7*(1–4). https://doi.org/10.1080/17588928.2015.1085375

Deschrijver, E., Bardi, L., Wiersema, J. R., & Brass, M. (2016). Behavioral measures of implicit theory of mind in adults with high functioning autism. *Cognitive Neuroscience*, *7*(1–4). https://doi.org/10.1080/17588928.2015.1085375

Deschrijver, E., Wiersema, J. R., & Brass, M. (2015). The interaction between felt touch and tactile consequences of observed actions: an action-based somatosensory congruency paradigm. *Social Cognitive and Affective Neuroscience*, *11*(7), 1162–1172. https://doi.org/10.1093/scan/nsv081

Deschrijver, E., Wiersema, J. R., & Brass, M. (2017). Action-based touch observation in high-functioning autism: Can compromised self-other distinction abilities link social and sensory problems in the autism spectrum? *Social Cognitive and Affective Neuroscience, 12 (2),* 273-282.

Deschrijver, E., Wiersema, J. R., & Brass, M. (2017). Disentangling Neural Sources of the Motor Interference Effect in High Functioning Autism: An EEG-Study. *Journal of Autism and Developmental Disorders*, *47*(3), 690–700. https://doi.org/10.1007/s10803-016-2991-2

Deschrijver, E., Wiersema, J. R., & Brass, M. (2017). The influence of action observation on action execution: Dissociating the contribution of action on perception, perception on action, and resolving conflict. *Cognitive, Affective and Behavioral Neuroscience*, *17*(2). https://doi.org/10.3758/s13415-016-0485-5

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). FMRI item analysis in a theory of mind task. *NeuroImage*, *55*(2), 705–712. https://doi.org/10.1016/j.neuroimage.2010.12.040

Döhnel, K., Schuwerk, T., Meinhardt, J., Sodian, B., Hajak, G., & Sommer, M. (2012). Functional activity of the right temporo-parietal junction and of the medial prefrontal cortex associated with true and false belief reasoning. *NeuroImage*, *60*(3), 1652–1661. https://doi.org/10.1016/j.neuroimage.2012.01.073

Döhnel, K., Schuwerk, T., Sodian, B., Hajak, G., Rupprecht, R., & Sommer, M. (2017). An fMRI study on the comparison of different types of false belief reasoning: False belief-based emotion and behavior attribution. *Social Neuroscience*, *12*(6), 730–742. https://doi.org/10.1080/17470919.2016.1241823

Dufour, N., Redcay, E., Young, L., Mavros, P. L., Moran, J. M., Triantafyllou, C., … Saxe, R. (2013). Similar Brain Activation during False Belief Tasks in a Large Sample of Adults with and without Autism. *PLoS ONE*, *8*(9). https://doi.org/10.1371/journal.pone.0075468

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., … Convit, A.

(2006). Introducing MASC: A movie for the assessment of social cognition. *Journal of Autism and Developmental Disorders*, *36*, 623–636. https://doi.org/10.1007/s10803-006-0107-0

El Kaddouri, R., Bardi, L., De Bremaeker, D., Brass, M., & Wiersema, J. R. (2019). Measuring spontaneous mentalizing with a ball detection task: putting the attention-check hypothesis by Phillips and colleagues (2015) to the test. *Psychological Research*, (123456789). https://doi.org/10.1007/s00426-019-01181-7

Fallon, N., Roberts, C., & Stancak, A. (2018). Functional networks of empathy: A systematic review and meta-analysis of fMRI studies of empathy for observed pain. *Preprint*.

Fan, Y.-T., Decety, J., Yang, C.-Y., Liu, J.-L., & Cheng, Y. (2010). Unbroken mirror neurons in autism spectrum disorders. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *51*, 981–988. https://doi.org/10.1111/j.1469-7610.2010.02269.x

Filmer, H. L., Fox, A., & Dux, P. E. (2019). Causal evidence of right temporal parietal junction involvement in implicit Theory of Mind processing. *NeuroImage*, *196*(February), 329–336. https://doi.org/10.1016/j.neuroimage.2019.04.032

Flynn, E., O'Malley, C., & Wood, D. (2004). A longitudinal, microgenetic study of the emergence of false belief understanding and inhibition skills. *Developmental Science*, *7*(1), 103–115. https://doi.org/10.1111/j.1467-7687.2004.00326.x

Gazzola, V., & Keysers, C. (2009). The observation and execution of actions share motor and somatosensory voxels in all tested subjects: Single-subject analyses of unsmoothed fMRI data. *Cerebral Cortex*, *19*, 1239–1255. https://doi.org/10.1093/cercor/bhn181

Gliga, T., Senju, A., Charman, T., & Johnson, M. H. (2014). Spontaneous Belief Attribution in Younger Siblings of Children on the Autism Spectrum, *50*(3), 903–913. https://doi.org/10.1037/a0034146

Grainger, S. A., Henry, J. D., Naughtin, C. K., Comino, M. S., & Dux, P. E. (2018). Implicit false belief tracking is preserved in late adulthood. *Quarterly Journal of Experimental Psychology*, 174702181773469. https://doi.org/10.1177/1747021817734690

Happé, F. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *843--855*, *66*(3).

Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour*, *61*(1), 139–151. https://doi.org/10.1006/anbe.2000.1518

Heyes, C. (2011). Automatic imitation. *Psychological Bulletin*, *137*(3), 463–483. https://doi.org/10.1037/a0022288

Heyes, C. (2014). Submentalizing : I Am Not Really Reading Your Mind. *Perspectives on Psychological Science*, *9*(2), 131–143. https://doi.org/10.1177/1745691613518076

Heyes, C. (2014). Submentalizing: I Am Not Really Reading Your Mind. *Perspectives on Psychological Science*, *9*, 131–143. https://doi.org/10.1177/1745691613518076

Hoffmann, F., Singer, T., & Steinbeis, N. (2015). Children's Increased Emotional Egocentricity Compared to Adults Is Mediated by Age-Related Differences in Conflict Processing. *Child Development*, *86*(3), 765–780. https://doi.org/10.1111/cdev.12338

Hogeveen, J., Obhi, S. S., Banissy, M. J., Santiesteban, I., Press, C., Catmur, C., & Bird, G. (2014). Task-dependent and distinct roles of the temporoparietal junction and inferior frontal cortex in the control of imitation. *Social Cognitive and Affective Neuroscience*, *10*(7), 1003–1009. https://doi.org/10.1093/scan/nsu148

Hughes, C. (1998). Executive function in preschoolers: Links with theory of mind and verbal ability. *British Journal of Developmental Psychology*, *16*(2), 233–253. https://doi.org/10.1111/j.2044-835X.1998.tb00921.x

Hyde, D. C., Betancourt, M. A., & Simon, C. E. (2015). Human Temporal-Parietal Junction Spontaneously Tracks Others ' Beliefs : A Functional Near-Infrared Spectroscopy

Study, *4846*, 4831–4846. https://doi.org/10.1002/hbm.22953
Iacoboni, M., & Dapretto, M. (2006). The mirror neuron system and the consequences of its dysfunction. *Nature Reviews. Neuroscience*, *7*(12), 942–51. https://doi.org/10.1038/nrn2024
Joseph, R. M., & Tager-Flusberg, H. (2004). The relationship of theory of mind and executive functions to symptom type and severity in children with autism. *Development and Psychopathology*, *16*(1), 137–155. https://doi.org/10.1017/S095457940404444X
Kana, R. K., Keller, T. A., Cherkassky, V. L., Minshew, N. J., & Just, M. A. (2009). Atypical frontal-posterior synchronization of Theory of Mind regions in autism during mental state attribution. *Social Neuroscience*, *4*(2), 135–152. https://doi.org/10.1080/17470910802198510.Atypical
Kana, R. K., Libero, L. E., Hu, C. P., Deshpande, H. D., & Colburn, J. S. (2014). Functional Brain Networks and White Matter Underlying Theory-of-Mind in Autism, *Social Cognitive and Affective Neuroscience, 9 (*1), 98-105. https://doi.org/10.1093/scan/nss106
Kana, R. K., Maximo, J. O., Williams, D. L., Keller, T. A., Schipul, S. E., Cherkassky, V. L., … Just, M. A. (2015). Aberrant functioning of the theory-of-mind network in children and adolescents with autism. *Molecular Autism*, 1–12. https://doi.org/10.1186/s13229-015-0052-x
Keysar, B., Barr, D. J., Balin, J. A., & Brauner, J. S. (2000). Taking perspective in conversation: The role of mutual knowledge in comprehension. *Psychological Science*, *11*(1), 32–38.
Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults, *89*, 25–41. https://doi.org/10.1016/S0010-0277(03)00064-7
Keysers, C., Kaas, J. H., & Gazzola, V. (2010). Somatosensation in social perception. *Nature Reviews. Neuroscience*, *11*, 417–428. https://doi.org/10.1038/nrn2919
Keysers, C., Wicker, B., Gazzola, V., Anton, J. L., Fogassi, L., & Gallese, V. (2004). A touching sight: SII/PV activation during the observation and experience of touch. *Neuron*, *42*, 335–346. https://doi.org/10.1016/S0896-6273(04)00156-4
Kliemann, D., Young, L., Scholz, J., & Saxe, R. (2008). The influence of prior record on moral judgment. *Neuropsychologia*, *46*(12), 2949–2957. https://doi.org/10.1016/j.neuropsychologia.2008.06.010.The
Koster-Hale, J., Bedny, M., & 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. https://doi.org/10.1016/j.cognition.2014.04.006
Koster-hale, J., Richardson, H., Velez, N., Asaba, M., Young, L., & Saxe, R. (2017). Mentalizing regions represent distributed , continuous, and abstract dimensions of others' beliefs. *NeuroImage*, *161*(August), 9–18. https://doi.org/10.1016/j.neuroimage.2017.08.026
Koster-hale, J., Saxe, R., Dungan, J., & 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). https://doi.org/10.1073/pnas.1207992110
Kovács, Á. M., Kühn, S., Gergely, G., Csibra, G., & Brass, M. (2014). Are all beliefs equal ? Implicit belief attributions recruiting core brain regions of Theory of Mind. *PLoS ONE*, *9*(9), e106558. https://doi.org/10.1371/journal.pone.0106558
Kovács, Á. M., Téglás, E., & Endress, A. D. (2010). The social sense: susceptibility to others' beliefs in human infants and adults. *Science (New York, N.Y.)*, *330*(2010), 1830–1834. https://doi.org/10.1126/science.1190792
Kronbichler, L., Tschernegg, M., Martin, A. I., Schurz, M., & Kronbichler, M. (2017).

Abnormal Brain Activation During Theory of Mind Tasks in Schizophrenia : A Meta-Analysis, *43*(6), 1240–1250. https://doi.org/10.1093/schbul/sbx073
Kulke, L., Duhn, B. Von, Schneider, D., & Rakoczy, H. (2018). Is Implicit Theory of Mind a Real and Robust Phenomenon ? Results From a Systematic Replication Study. https://doi.org/10.1177/0956797617747090
Kulke, L., & Göttingen, G. (2017). How robust are anticipatory looking measures of Theory of Mind ? Replication attempts across the life span. *Cognitive Development*, (October), 0–1. https://doi.org/10.1016/j.cogdev.2017.09.001
Kulke, L., Johannsen, J., & Rakoczy, H. (2019). Why can some implicit Theory of Mind tasks be replicated and others cannot? A test of mentalizing versus submentalizing accounts. *PLoS ONE*, *14*(3). https://doi.org/10.1371/journal.pone.0213772
Lamm, C., Bukowski, H., & Silani, G. (2016). From shared to distinct self–other representations in empathy.*Philosophical Transactions of the Royal Society B: Biological Sciences, 371* (1686).
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), 2492–502. https://doi.org/10.1016/j.neuroimage.2010.10.014
Lee, J., Horan, W. P., Wynn, J. K., & Green, M. F. (2016). Neural Correlates of Belief and Emotion Attribution in Schizophrenia, 1–13. https://doi.org/10.1371/journal.pone.0165546
Lee, L., Harkness, K. L., Sabbagh, M. A., & Jacobson, J. A. (2005). Mental state decoding abilities in clinical depression. *Journal of Affective Disorders*, *86*(2–3), 247–258. https://doi.org/10.1016/j.jad.2005.02.007
Leekam, S. R., & Perner, J. (1991). Does the autistic child have a metarepresentational deficit? *Cognition*, *40*(3), 203–218. Retrieved from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed2&NEWS=N&AN=1786675
Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in " theory of mind ," *8*(12). https://doi.org/10.1016/j.tics.2004.10.001
Leslie, A. M., & Thaiss, L. (1992). Domain specificity in conceptual development: neuropsychological evidence from autism. *Cognition*, *43*(3), 225–251. Retrieved from http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed2&NEWS=N&AN=1643814
Lombardo, M. V., & Baron-Cohen, S. (2011). The role of the self in mindblindness in autism. *Consciousness and Cognition*, *20*(1), 130–140. https://doi.org/10.1016/j.concog.2010.09.006
Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., & Baron-Cohen, S. (2011). Specialization of right temporo-parietal junction for mentalizing and its relation to social impairments in autism. *NeuroImage*, *56*(3), 1832–1838. https://doi.org/10.1016/j.neuroimage.2011.02.067
Lombardo, M. V., Chakrabarti, B., Bullmore, E. T., Wheelwright, S. J., Sadek, S. a., Suckling, J., … Baron-Cohen, S. (2007). Shared neural circuits for mentalizing about the self and others, 1–10.
Marticorena, D. C. W., Ruiz, A. M., Mukerji, C., Goddu, A., & Santos, L. R. (2011). Monkeys represent others' knowledge but not their beliefs. *Developmental Science*, *14*(6), 1406–1416. https://doi.org/10.1111/j.1467-7687.2011.01085.x
Martin, A., & Santos, L. R. (2014). The Origins of Belief Representation: Monkeys Fail to Automatically Represent Others' Beliefs. *Cognition*, *130*(3), 300–308. https://doi.org/10.1016/j.cognition.2013.11.016.The
Martin, A., & Santos, L. R. (2016). What Cognitive Representations Support Primate

Theory of Mind? *Trends in Cognitive Sciences*, *20*(5), 375–382. https://doi.org/10.1016/j.tics.2016.03.005

Martineau, J., Andersson, F., Barthélémy, C., Cottier, J. P., & Destrieux, C. (2010). Atypical activation of the mirror neuron system during perception of hand motion in autism. *Brain Research*, *1320*, 168–175. https://doi.org/10.1016/j.brainres.2010.01.035

Mason, R. A., Williams, D. L., Kana, R. K., Minshew, N., & Just, M. A. (2008). Theory of Mind disruption and recruitment of the right hemisphere during narrative comprehension in autism. *Neuropsychologia*, *46*(1), 269–280. https://doi.org/10.1016/j.neuropsychologia.2007.07.018

Mitchell, J. P. (2009). Inferences about mental states. *Philosophical Transactions of the Royal Society B: Biological Sciences, 364*(1521), 1309-1316. 1309–1316. https://doi.org/10.1098/rstb.2008.0318

Murdaugh, D. L., Nadendla, K. D., & Kana, R. K. (2014). Differential role of temporoparietal junction and medial prefrontal cortex in causal inference in autism : An independent component analysis. *Neuroscience Letters*, *568*, 50–55. https://doi.org/10.1016/j.neulet.2014.03.051

Naughtin, C. K., Horne, K., Schneider, D., Venini, D., York, A., & Dux, P. E. (2017a). Do Implicit and Explicit Belief Processing Share Neural Substrates ?, *4772*, 4760–4772. https://doi.org/10.1002/hbm.23700

Naughtin, C. K., Horne, K., Schneider, D., Venini, D., York, A., & Dux, P. E. (2017b). Do implicit and explicit belief processing share neural substrates? *Human Brain Mapping*, *38*(9), 4760–4772. https://doi.org/10.1002/hbm.23700

Nijhof, A., Brass, M., Bardi, L., & Wiersema, J. R. (2016). Measuring Mentalizing Ability: A Within-Subjects Comparison between an Explicit and Implicit Version of a Ball Detection Task. *PLoS ONE*, *11*(10).

Nijhof, A. D., Bardi, L., Brass, M., & Wiersema, J. R. (2018). Brain activity for spontaneous and explicit mentalizing in adults with autism spectrum disorder : An fMRI study, (February). https://doi.org/10.1016/j.nicl.2018.02.016

Nijhof, A. D., Brass, M., & Wiersema, J. R. (2017). Spontaneous mentalizing in neurotypicals scoring high versus low on symptomatology of autism spectrum disorder. *Psychiatry Research*, *258*(August), 15–20. https://doi.org/10.1016/j.psychres.2017.09.060

Nilsson, K. K., & de López, K. J. (2016). Theory of Mind in Children With Specific Language Impairment: A Systematic Review and Meta-Analysis. *Child Development*, *87*(1), 143–153. https://doi.org/10.1111/cdev.12462

Nobusako, S., Nishi, Y., Nishi, Y., Shuto, T., & Asano, D. (2017). Transcranial Direct Current Stimulation of the Temporoparietal Junction and Inferior Frontal Cortex Improves Imitation-Inhibition and Perspective-Taking with no Effect on the Autism-Spectrum Quotient Score, *11*(May), 1–12. https://doi.org/10.3389/fnbeh.2017.00084

Oberman, L. M., Ramachandran, V. S., & Pineda, J. a. (2008). Modulation of mu suppression in children with autism spectrum disorders in response to familiar or unfamiliar stimuli: The mirror neuron hypothesis. *Neuropsychologia*, *46*, 1558–1565. https://doi.org/10.1016/j.neuropsychologia.2008.01.010

Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science (New York, N.Y.)*, *308*(2005), 255–258. https://doi.org/10.1126/science.1107621

Overwalle, F. Van. (2009). Social Cognition and the Brain : A Meta-Analysis, *858*, 829–858. https://doi.org/10.1002/hbm.20547

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

https://doi.org/10.1016/j.neuroimage.2009.06.009
Özdem, C., Brass, M., Schippers, A., Van der Cruyssen, L., & Van Overwalle, F. (2019). The neural representation of mental beliefs held by two agents. *Cognitive, Affective and Behavioral Neuroscience*, *19*(6), 1433–1443. https://doi.org/10.3758/s13415-019-00714-2
Özdem, C., Brass, M., Van der Cruyssen, L., & Van Overwalle, F. (2017). The overlap between false belief and spatial reorientation in the temporo-parietal junction: The role of input modality and task. *Social Neuroscience*, *12*(2), 207–217. https://doi.org/10.1080/17470919.2016.1143027
Ozonoff, S., Pennington, B. F., & Rogers, S. J. (1991). Executive Function Deficits in High-Functioning Autistic Individuals: Relationship to Theory of Mind. *Journal of Child Psychology and Psychiatry*, *32*(7), 1081–1105. https://doi.org/10.1111/j.1469-7610.1991.tb00351.x
Parkinson, C., Kleinbaum, A. M., & Wheatley, T. (2018). Similar neural responses predict friendship. *Nature Communications*, *9*(1). https://doi.org/10.1038/s41467-017-02722-7
Pellicano, E. (2007). Links Between Theory of Mind and Executive Function in Young Children With Autism: Clues to Developmental Primacy. *Developmental Psychology*, *43*(4), 974–990. https://doi.org/10.1037/0012-1649.43.4.974
Perner, J., Aichhorn, M., Kronbichler, M., Staffen, W., Perner, J., Aichhorn, M., … Ladurner, G. (2007). Thinking of mental and other representations : The roles of left and right temporo-parietal junction Thinking of mental and other representations : The roles of left and right temporo-parietal junction, *919*. https://doi.org/10.1080/17470910600989896
Perner, J., Leekam, S., & Leekam, S. (2008). The curious incident of the photo that was accused of being false : Issues of domain specificity in development , autism , and brain imaging The curious incident of the photo that was accused of being false : Issues of domain specificity in development , , *218*. https://doi.org/10.1080/17470210701508756
Peterson, C. C., & Slaughter, V. P. (2007). Social maturity and theory of mind in typically developing children and those on the autism spectrum. *Journal of Child Psychology and Psychiatry*, *48*(12), 1243–1250.
Phillips, J., Desmond, C., O., Surtees, A. D. R., Xin, Y., Williams, S., Saxe, R., & Frank, M. C. (2015). A second look at automatic theory of mind: Reconsidering Kovacs, Teglas, and Endress (2010). *Psychological Science*, *26*(9), 1353–1367.
Phillips, J., & Norby, A. (2019). Factive theory of mind. *Mind and Language*, (February), 1–24. https://doi.org/10.1111/mila.12267
Powell, L. J., Hobbs, K., Bardis, A., Carey, S., & Saxe, R. (2018). Replications of implicit theory of mind tasks with varying representational demands. *Cognitive Development*, *46*(October 2017), 40–50. https://doi.org/10.1016/j.cogdev.2017.10.004
Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526. https://doi.org/10.1017/S0140525X00076512
Priewasser, B., Rafetseder, E., Gargitter, C., & Perner, J. (2018). Helping as an early indicator of a theory of mind: Mentalism or Teleology? *Cognitive Development*, *46*(February 2017), 69–78. https://doi.org/10.1016/j.cogdev.2017.08.002
Qureshi, A. W., Apperly, I. A., & Samson, D. (2010). Executive function is necessary for perspective selection, not Level-1 visual perspective calculation: Evidence from a dual-task study of adults. *Cognition*, *117*(2), 230–236. https://doi.org/10.1016/j.cognition.2010.08.003
Ritchie, G. (2009). Variants of Incongruity Resolution. *Journal of Literary Theory*, *2*, 1–20. https://doi.org/10.1515/JLT.2009.015

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, *27*, 169–192. https://doi.org/10.1146/annurev.neuro.27.070203.144230
Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews. Neuroscience*, *2*(9), 661–70. https://doi.org/10.1038/35090060
Robalino, N., & Robson, A. (2012). The economic approach to "theory of mind." *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2224–2233. https://doi.org/10.1098/rstb.2012.0124
Rothmayr, C., Sodian, B., Hajak, G., Döhnel, K., Meinhardt, J., & Sommer, M. (2011a). Common and distinct neural networks for false-belief reasoning and inhibitory control. *NeuroImage*, *56*(3), 1705–1713. https://doi.org/10.1016/j.neuroimage.2010.12.052
Rothmayr, C., Sodian, B., Hajak, G., Döhnel, K., Meinhardt, J., & Sommer, M. (2011b). Common and distinct neural networks for false-belief reasoning and inhibitory control. *NeuroImage*, *56*(3), 1705–1713. https://doi.org/10.1016/j.neuroimage.2010.12.052
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.
Russell, J., Saltmarsh, R., & Hill, E. (1999). What do executive factors contribute to the failure on false belief tasks by children with autism? *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *40*(6), 859–868. https://doi.org/10.1017/S0021963099004229
Samson, A. C., Hempelmann, C. F., Huber, O., & Zysset, S. (2009). Neural substrates of incongruity-resolution and nonsense humor. *Neuropsychologia*, *47*(4), 1023–1033. https://doi.org/10.1016/j.neuropsychologia.2008.10.028
Samson, A. C., Zysset, S., Huber, O., Samson, A. C., Zysset, S., Cognitive, O. H., … Huber, O. (2008). Cognitive humor processing: Different logical mechanisms in nonverbal cartoons—an fMRI study. *Social Neuroscience*, *3*(2), 125–140. https://doi.org/10.1080/17470910701745858
Samson, D. (2009). Reading other people's mind: Insights from neuropsychology. *Journal of Neuropsychology*, *3*, 3–16. https://doi.org/10.1348/174866408X377883
Samson, D., Apperly, I. A., Chiavarino, C., & Humphreys, G. W. (2004). Left temporoparietal junction is necessary for representing someone else ´ s belief, *7*(5), 499–500. https://doi.org/10.1038/nn1223
Samson, D., Apperly, I. A., Kathirgamanathan, U., & Humphreys, G. W. (2005). Seeing it my way : a case of a selective deficit in inhibiting self-perspective, 1102–1111. https://doi.org/10.1093/brain/awh464
Santiesteban, I., Banissy, M. J., Catmur, C., & Bird, G. (2012). Enhancing social ability by stimulating right temporoparietal junction. *Current Biology*, *22*(23), 2274–2277. https://doi.org/10.1016/j.cub.2012.10.018
Santiesteban, I., Banissy, M. J., Catmur, C., & Bird, G. (2015). Functional lateralization of temporoparietal junction – imitation inhibition, visual perspective-taking and theory of mind. *European Journal of Neuroscience*, *42*(8), 2527–2533.
Santiesteban, I., White, S., Cook, J., Gilbert, S. J., Heyes, C., & Bird, G. (2012). Training social cognition: From imitation to Theory of Mind. *Cognition*, *122*(2011), 228–235. https://doi.org/10.1016/j.cognition.2011.11.004
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. https://doi.org/10.1016/S1053-8119(03)00230-1
Saxe, R., Moran, J. M., Scholz, J., & Gabrieli, J. (2006). Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *Social Cognitive and Affective Neuroscience*, *1*, 229–234. https://doi.org/10.1093/scan/nsl034

Saxe, R., & Powell, L. J. (2006). It's the Thought That Counts. *Psychological Science*, *17*(8), 692–700.
Saxe, R., Schulz, L. E., & Jiang, Y. V. (2006). Reading minds versus following rules : Dissociating theory of mind and executive control in the brain, *1*, 284–298. https://doi.org/10.1080/17470910601000446
Saxe, R., & Wexler, A. (2005a). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia*. https://doi.org/10.1016/j.neuropsychologia.2005.02.013
Saxe, R., & Wexler, A. (2005b). Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia*, *43*(10), 1391–9. https://doi.org/10.1016/j.neuropsychologia.2005.02.013
Schaafsma, S. M., Pfaff, D. W., Spunt, R. P., & Adolphs, R. (2016). Deconstructing and Reconstructing Theory of Mind. *Trends in Cognitive Sciences*, *19*(2), 65–72. https://doi.org/10.1016/j.tics.2014.11.007.Deconstructing
Scheeren, A. M., De Rosnay, M., Koot, H. M., & Begeer, S. (2013). Rethinking theory of mind in high-functioning autism spectrum disorder. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *54*, 628–635. https://doi.org/10.1111/jcpp.12007
Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *The Behavioral and Brain Sciences*, *36*, 393–414. https://doi.org/10.1017/S0140525X12000660
Schneider, D., Bayliss, A. P., Becker, S. I., & Dux, P. E. (2012). Eye movements reveal sustained implicit processing of others' mental states. *Journal of Experimental Psychology: General*, *141*(3), 433–438. https://doi.org/10.1037/a0025458
Schneider, D., Lam, R., Bayliss, A. P., & Dux, P. E. (2012). Cognitive Load Disrupts Implicit Theory-of-Mind Processing. *Psychological Science*, *23*(8), 842–847. https://doi.org/10.1177/0956797612439070
Schneider, D., Nott, Z. E., & Dux, P. E. (2014). Task instructions and implicit theory of mind. *Cognition*, *133*(1), 43–47. https://doi.org/10.1016/j.cognition.2014.05.016
Schneider, D., Slaughter, V. P., Bayliss, A. P., & Dux, P. E. (2013). A temporally sustained implicit theory of mind deficit in autism spectrum disorders. *Cognition*, *129*(2), 410–417. https://doi.org/10.1016/j.cognition.2013.08.004
Schneider, D., Slaughter, V. P., Becker, S. I., & Dux, P. E. (2014). Implicit false-belief processing in the human brain. *NeuroImage*, *101*, 268–275. https://doi.org/10.1016/j.neuroimage.2014.07.014
Schneider, D., Slaughter, V. P., & Dux, P. E. (2017). Current evidence for automatic Theory of Mind processing in adults. *Cognition*, *162*, 27–31. https://doi.org/10.1016/j.cognition.2017.01.018
Schurz, M., Aichhorn, M., Martin, A., & Perner, J. (2013). Common brain areas engaged in false belief reasoning and visual perspective taking : a meta-analysis of functional brain imaging studies. *Frontiers in Human Neuroscience*, *7*(November), 1–14. https://doi.org/10.3389/fnhum.2013.00712
Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind : A meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, *42*, 9–34. https://doi.org/10.1016/j.neubiorev.2014.01.009
Schurz, M., Tholen, M. G., Perner, J., Mars, R. B., & Sallet, J. (2017). Specifying the Brain Anatomy Underlying Temporo-Parietal Junction Activations for Theory of Mind : A Review using Probabilistic Atlases from Different Imaging Modalities, *4805*(June), 4788–4805. https://doi.org/10.1002/hbm.23675
Schuwerk, T., Jarvers, I., Vuori, M., & Sodian, B. (2016). Implicit Mentalizing Persists beyond Early Childhood and Is Profoundly Impaired in Children with Autism Spectrum

Condition, *7*(October), 1–9. https://doi.org/10.3389/fpsyg.2016.01696
Schuwerk, T., Vuori, M., & Sodian, B. (2015). Implicit and explicit Theory of Mind reasoning in autism spectrum disorders : The impact of experience. https://doi.org/10.1177/1362361314526004
Scott, R. M., & Baillargeon, R. (2017). Early False-Belief Understanding. *Trends in Cognitive Sciences*, *21*(4), 237–249. https://doi.org/10.1016/j.tics.2017.01.012
Scott, R. M., Baillargeon, R., Song, H.-J., & Leslie, A. M. (2010). Attributing false beliefs about non-obvious properties at 18 months. *Cognitive Psychology*, *61*(4), 366–395. https://doi.org/10.1016/j.cogpsych.2010.09.001.Attributing
Senju, A. (2013a). Atypical development of spontaneous social cognition in autism spectrum disorders. *Brain and Development*, *35*(2), 96–101. https://doi.org/10.1016/j.braindev.2012.08.002
Senju, A. (2013b). Spontaneous theory of mind and its absence in autism spectrum disorders, *18*(2), 108–113. https://doi.org/10.1177/1073858410397208.Spontaneous
Senju, A., Southgate, V., White, S., & Frith, U. (2009). Mindblind eyes: an absence of spontaneous theory of mind in Asperger syndrome. *Science (New York, N.Y.)*, *325*(2009), 883–885. https://doi.org/10.1126/science.1176170
Silani, G., Lamm, C., Ruff, C. C., & Singer, T. (2013). Right supramarginal gyrus is crucial to overcome emotional egocentricity bias in social judgments. *Journal of Neuroscience*, *33*(39), 15466–15476. https://doi.org/10.1523/JNEUROSCI.1488-13.2013
Sodian, B., & Thoermer, C. (2008). Precursors to a theory of mind in infancy: Perspectives for research on autism. *Quarterly Journal of Experimental Psychology*, *61*(1), 27–39. https://doi.org/10.1080/17470210701508681
Sommer, M., Döhnel, K., Jarvers, I., Blaas, L., Singer, M., Nöth, V., … Rupprecht, R. (2018). False belief reasoning in adults with and without autistic spectrum disorder: Similarities and differences. *Frontiers in Psychology*, *9*(FEB), 1–12. https://doi.org/10.3389/fpsyg.2018.00183
Sommer, M., Döhnel, K., Sodian, B., Meinhardt, J., Thoermer, C., & Hajak, G. (2007). Neural correlates of true and false belief reasoning, *35*, 1378–1384. https://doi.org/10.1016/j.neuroimage.2007.01.042
Southgate, V., & Hamilton, A. F. D. C. (2008). Unbroken mirrors: challenging a theory of Autism. *Trends in Cognitive Sciences*, *12*, 225–229. https://doi.org/10.1016/j.tics.2008.03.005
Sowden, S., & Catmur, C. (2013). The role of the right temporoparietal junction in the control of imitation. *Cerebral Cortex*, (2010), 1–7. https://doi.org/10.1093/cercor/bht306
Sowden, S., & Catmur, C. (2015). The role of the right temporoparietal junction in the control of imitation. *Cerebral Cortex*, *25*(4), 1107–1113. https://doi.org/10.1093/cercor/bht306
Sowden, S., Koehne, S., Catmur, C., Dziobek, I., & Bird, G. (2015). Intact Automatic Imitation and Typical Spatial Compatibility in Autism Spectrum Disorder: Challenging the Broken Mirror Theory. *Autism Research*. https://doi.org/10.1002/aur.1511
Sowden, S., & Shah, P. (2014). Self-other control: a candidate mechanism for social cognitive function. *Frontiers in Human Neuroscience*, *8*, 789. https://doi.org/10.3389/fnhum.2014.00789
Sowden, S., Wright, G. R. T., Banissy, M. J., & Catmur, C. (2015). Transcranial Current Stimulation of the Temporoparietal Junction Improves Lie Detection Report Transcranial Current Stimulation of the Temporoparietal Junction Improves Lie Detection. *Current Biology*, *25*(18), 2447–2451. https://doi.org/10.1016/j.cub.2015.08.014

Spengler, S., Bird, G., & Brass, M. (2010). Hyperimitation of actions is related to reduced understanding of others' minds in autism spectrum conditions. *Biological Psychiatry*, *68*, 1148–1155. https://doi.org/10.1016/j.biopsych.2010.09.017

Spengler, S., von Cramon, D. Y., & Brass, M. (2010). Resisting motor mimicry: Control of imitation involves processes central to social cognition in patients with frontal and temporo-parietal lesions. *Social Neuroscience*, *5*(4), 401–416. https://doi.org/10.1080/17470911003687905

Spengler, S., Von Cramon, D. Y., & Brass, M. (2009). Control of shared representations relies on key processes involved in mental state attribution. *Human Brain Mapping*, *30*(June), 3704–3718. https://doi.org/10.1002/hbm.20800

Steinbeis, N., Bernhardt, B. C., & Singer, T. (2015). Age-related differences in function and structure of rSMG and reduced functional connectivity with DLPFC explains heightened emotional egocentricity bias in childhood. *Social Cognitive and Affective Neuroscience*, *10*(2), 302–310. https://doi.org/10.1093/scan/nsu057

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological Science*, *18*, 580–586. https://doi.org/10.1111/j.1467-9280.2007.01943.x

Thompson, J. R. (2014). Signature Limits in Mindreading Systems, *38*, 1432–1455. https://doi.org/10.1111/cogs.12117

van der Weiden, A., Prikken, M., & van Haren, N. E. M. (2015). Self–other integration and distinction in schizophrenia: A theoretical analysis and a review of the evidence. *Neuroscience & Biobehavioral Reviews*, *57*, 220–237. https://doi.org/10.1016/j.neubiorev.2015.09.004

von Mohr, M., Finotti, G., Ambroziak, K. B., & Tsakiris, M. (2019). Do you hear what I see? An audio-visual paradigm to assess emotional egocentricity bias. *Cognition and Emotion, 34*(4), 756-770*,* https://doi.org/10.1192/bjp.111.479.1009-a

Wang, L., & Leslie, A. M. (2016). Is Implicit Theory of Mind the "Real Deal"? The Own-Belief/True-Belief Default in Adults and Young Preschoolers. *Mind and Language*, *31*(2), 147–176. https://doi.org/10.1111/mila.12099

White, S. J., Frith, U., Rellecke, J., Al-noor, Z., & Gilbert, S. J. (2014). Autistic adolescents showatypical activation of the brain0s mentalizing system even without a prior history of mentalizing problems. *Neuropsychologia*, *56*, 17–25. https://doi.org/10.1016/j.neuropsychologia.2013.12.013

Wiesmann, C. G., Friederici, A. D., Singer, T., & Steinbeis, N. (2017). Implicit and explicit false belief development in preschool children, *Developmental Science, 20* (5), 1–15. https://doi.org/10.1111/desc.12445

Williams, J. H. G., Whiten, A., & Singh, T. (2004). A systematic review of action imitation in autistic spectrum disorder. *Journal of Autism and Developmental Disorders*, *34*(3), 285–299. https://doi.org/10.1023/B:JADD.0000029551.56735.3a

Wysocka, J., Golec, K., Haman, M., Wolak, T., Kochański, B., & Pluta, A. (2020). Processing False Beliefs in Preschool Children and Adults: Developing a Set of Custom Tasks to Test the Theory of Mind in Neuroimaging and Behavioral Research. *Frontiers in Human Neuroscience*, *14*(April), 1–15. https://doi.org/10.3389/fnhum.2020.00119

Young, L., Albert, J., Hauser, M., Pascual-leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. https://doi.org/10.1073/pnas.0914826107

Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment, *104*(20), 8235–8240.

Young, L., & Saxe, R. (2009). Innocent intentions- A correlation between forgiveness for accidental harm and neural activity. *Neuropsychologia*, *47*(10), 2065–2072.

Zaitchik, D. (1990). When representations conflict with reality: The preschooler's problem with false beliefs and "false" photographs. *Cognition*, *35*(1), 41–68. https://doi.org/10.1016/0010-0277(90)90036-J

Zelazo, P. D., Jacques, S., Burack, J. A., & Frye, D. (2002). The Relation between Theory of Mind and Rule Use: Evidence from Persons with Autism-Spectrum Disorders. *Infant and Child Development*, *11*, 171–195. https://doi.org/10.1002/icd

Zwickel, J., White, S. J., Coniston, D., Senju, A., & Frith, U. (2011). Exploring the building blocks of social cognition: Spontaneous agency perception and visual perspective taking in autism. *Social Cognitive and Affective Neuroscience*, *6*, 564–571. https://doi.org/10.1093/scan/nsq088