TOWARDS A NEUROCOGNITIVE THEORY OF MIND: HOW CONTROL AND REASONING PROCESSES CONTRIBUTE TO ADULT MENTALIZING

by

CHARLOTTE EMILY HARTWRIGHT

A thesis submitted to

The University of Birmingham

for the degree of

DOCTOR OF PHILOSOPHY

School of Psychology

College of Life and Environmental Sciences

University of Birmingham

October 2013

ABSTRACT

A series of neuroimaging experiments were conducted using adult participants to explore the neurocognitive bases of reasoning and control processes in Theory of Mind (ToM). Through careful manipulation of psychologically relevant parameters, these were designed to modulate neural regions considered important for ToM, the temporoparietal junction (TPJ) and medial prefrontal cortex (mPFC), alongside regions more typically associated with executive function, the ventrolateral and dorsomedial prefrontal cortices (vlPFC / dmPFC respectively). This enabled close inspection of the functional profile of these regions, in the context of mentalizing. The core findings were: 1) TPJ was modulated by the valence of mental states. Thus, TPJ does not simply respond to mental representation; the content of the representation is important. 2) The presence of activation in rostral mPFC was manipulated by varying the mode of reasoning. Context is therefore relevant to how adults approach ToM. 3) A neural dissociation was identified between two accounts of control processes for ToM in vlPFC and dmPFC. Such processes mediate the expression of certain ToM concepts. Together, these findings suggest that a neurocognitive account of ToM should describe a flexible system which adapts to the specific conceptual and contextual demands of the social world at that time.

~ FOR BERTIE ~

with fond memories of the Brain Club

ACKNOWLEDGEMENTS

Firstly, I would like to thank my PhD supervisors, Professor Ian Apperly and Dr Peter Hansen. Thank you for many hours of academic mentoring and scientific instruction. Ian, thank you for the thought provoking exchange over the years; your ability to fractionate complex psychological phenomena is incredible and has provided much inspiration. Peter, your technical instruction has been absolutely first class. Thank you for giving me such a comprehensive background in imaging. Thank you also for introducing me to a lasting friend; I will forever enjoy the benefits of being able to streamline my work processes with Bash.

Thank you also to my dear friends Blaire, Chloé, Emily, Magda and Sam. Thank you for listening to me prattle on about ill-structured syntax, faulty chillers and excessive head motion. The coffee might have always been good, but I suspect at times my company wasn't!

Thank you to my family. You all have done so much to support me during my studies. Thank you, Nan, for your endless enthusiasm for my academic pursuits. Mum, thank you for the many hours of first class grand parenting; for getting out the 'ticka tocka' whilst I quietly shuffle off to the study. Thank you, Dad, for countless roasts with those monolithic Yorkshires – a welcome break from many hours spent writing scripts or analyzing data.

Lastly, thank you to Rod. Without you, I would never have had the confidence to embark on this journey, instead, forever feeling unfulfilled by corporate drudgery. Thank you for always believing in me, for providing encouragement and eternal positivity. Thank you for being a support every day; for being the (pilot) pilot participant, the peer reviewer and the IT helpdesk, whilst remaining my wonderful, loving husband.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1:

GENERAL INTRODUCTION

"When you are a Bear of Very Little Brain, and you Think of Things,

you find sometimes that a Thing which seemed very Thingish inside you

is quite different when it gets out into the open and has other people looking at it."

A. A. Milne, The House at Pooh Corner (1928)

INTRODUCTION

It often feels that we navigate the social world with ease, using our social cognition to rationalise and predict the observable, and unobservable, causes of behaviour. Psychologists and Neuroscientists have, however, demonstrated that this is not the case. Negotiating the social world is cognitively and neurally effortful; when we look inside 'the black box', the brain is in a constant state of social chatter (Schilbach, Eickhoff, Rotarska-Jagiela, Fink, & Vogeley, 2008), suspending such processes only briefly to make way for others. Young children, for example, are notoriously bad at modelling the minds of others (Wellman, Cross, & Watson, 2001). When it comes to being a know-it-all, the pre-schooler wins first prize, where bias towards new information replaces their knowledge of what they knew before – *"but I knew it all along, Mummy"* (Birch & Bloom, 2004; Gopnik & Astington, 1988). In adulthood, we improve, yet still our brains tend to reference our own perspective, when trying to take someone else's. We even automatically compute other people's viewpoints, when we need not (Epley, Morewedge, & Keysar, 2004; Ramsey, Hansen, Apperly, & Samson, 2013). To make decisions about what is going on in someone else's mind, we call on parts of the brain involved in autobiographical memory, prospection and storytelling (Mar, 2011; Spreng, Mar, & Kim, 2009); we bring together creative streams of our past and future selves, to service the current social referent. For similar reasons, our brains automatically engage neural

regions that are involved in the execution of our own movement, when viewing the movements of someone else (Pobric & Hamilton, 2006; Ramsey & Hamilton, 2012). Taken together, as Winnie the Pooh suggests, how 'Things' feel inside can be very different to what is there when opened up for all to see: our 'social qualia' may not reflect our social cognition.

As the fields of Psychology and Neuroscience come together, it is now becoming possible to see the hard work that goes into negotiating the social word. Perhaps unlike most other research fields, however, those involved in understanding social cognition face a unique challenge with ToM in that, whilst the vast majority of us have *it*, and though most of us use *it* every day, 'Theory of Mind' is a term that is largely unfamiliar outside of the academic domain. Like the social beings in which *it* is housed, ToM encompasses myriad abilities and processes, which ultimately serve the capacity to know what *it* is all about in another person's mind. Examining this essentially opaque ability, however, requires research which seeks to either constrain ToM into a unified concept, or attempts to systematically vary potential constituent parts to better describe what *it* really is and, importantly, how *it* is done. In examining ToM in typically functioning adults, my thesis adopts the latter approach.

This thesis presents four studies of adult Theory of Mind (ToM). Using functional Magnetic Resonance Imaging (fMRI) and Transcranial Magnetic Stimulation (TMS), this thesis outlines four novel experiments which explore the functional profile of two core mentalizing regions – the temporoparietal junction (TPJ) and the medial prefrontal cortex (mPFC) – alongside two regions believed to provide some of the executive control processes required to facilitate the expression of ToM – the ventrolateral and dorsal medial prefrontal cortices (vlPFC / dmPFC, respectively). This first chapter provides an historical account of ToM. The overview given is biased towards the themes and protocols that feature throughout this thesis. As such, it should not to be considered an exhaustive or even complete account of

how ToM research has evolved. However, the intention is to ground the reader in the research context in which this thesis is based.

The Role of Misinformation in Representation

The term 'Theory of Mind' was coined in the late 1970's in a seminal paper by Premack and Woodruff (1978). In a series of tasks, the authors sought to determine whether Sarah, a 14 year old chimpanzee, could represent the mind state of a human agent. Sarah was shown video footage which depicted a human actor in a cage struggling to retrieve a banana, which was out of reach. After watching the footage, Sarah was required to choose from a selection of objects, one of which could assist the actor in retrieving the banana, such as a rod to extend the actor's reach. Remarkably, Sarah typically selected the appropriate object, which the authors argue demonstrated that she could impute both the actor's intentional state – to retrieve the banana – and the actor's knowledge state – that the actor had the capacity to know how to retrieve the banana using the object. In this context, Premack and Woodruff (1978) described how the ability to think in such a way reflects a 'theory' regarding the contents of the actor's mind: a 'theory' as these mental phenomena are unobservable entities which can be used to generate predictions.

The study by Premack and Woodruff (1978) caused a flurry of commentaries, of which some have informed the approach to studying ToM in the present thesis. In particular, one critique, framed with reference to children's delight at Punch and Judy shows, suggested that ToM would be more believable if demonstrated, first, in the absence of overt training – as it was difficult to extend their findings to the wider chimpanzee population – and, second, where some form of outward expression, made by the subject, reflected a behavioural prediction referenced to an agent acting on misinformation. In the case of Punch and Judy,

this is children's joy with Punch as he is about to throw the box, which he falsely believes contains Judy, over the cliff (Dennett, 1978). The expression, in this case the children's glee, demonstrates that they expected Punch to throw the box off the cliff to serve his own desire, that is, to cause the demise of poor Judy. This approach was formalised into a false belief task (Baron-Cohen, Leslie, & Frith, 1985; Wimmer & Perner, 1983) which has heavily featured in ToM research and the current thesis.

In a typical false belief task, a character, Maxi, places his chocolate in cupboard X. He leaves the room and, in his absence, his mother moves the chocolate into cupboard Y. Participants are then asked to indicate which cupboard Maxi will look in to find his chocolate when he returns. The overall premise of this scenario is that, in order to identify that Maxi will look in cupboard X, one's prediction must be constrained by Maxi's beliefs. One cannot meaningfully refer to anything other than the unobservable contents of Maxi's mind. In their initial study involving 4- to 9-year old children, Wimmer and Perner (1983) identified a developmental trend, where children under the age of 5 years tended to (incorrectly) point towards cupboard Y – the real location of the chocolate – whereas older children correctly identified that Maxi would believe that his chocolate was in cupboard X. Interestingly, none of the children selected the control location, suggesting that they were either systematically choosing the location that complemented their own knowledge state – as was the case with younger children – or using a theory of Maxi's mind to successfully predict his behaviour. Using the format of this 'unexpected transfer task', Baron-Cohen et al. (1985) demonstrated that autistic children behaved very much like the younger children described by Wimmer and Perner (1983). Unlike developmentally typical controls and children with Down's syndrome, children with autism systematically predicted the behaviour of an agent on the basis of their own perspective. The authors conclude that this reflects a cognitive deficit in the ability to

ascribe higher order mental states, which has also been later speculated to be present in clinical disorders such as schizophrenia (Brune, 2005), bipolar affective disorder (Kerr, Dunbar, & Bentall, 2003) and antisocial personality disorder (Dolan & Fullam, 2004).

In order to better understand the basis of young children's difficulty with false belief reasoning, a control task was devised which negates the need for representing mental content but, nonetheless, requires representation of an alternative, non-mental referent (Zaitchik, 1990). As illustrated in the following overview, the 'false photograph' task adopts a similar format to the unexpected transfer false belief task, yet is solved without using ToM.

Belief format

- Actor A places object into location X then leaves
- Actor B moves the object to location Y
- Test question: "Where does A *think* the object is?"

Photo format

- Actor A places object into location X then takes a photo of it
- Actor B moves the object to location Y
- Test question: "*In the picture*, where is the object?"

False Belief / Photo Task Overview

Zaitchik (1990)

The two tasks comprise a similar structure, although one requires reference to the content of Actor A's mind, whereas the other refers to the content of a physical representation. The photo is kept from view from the participant, thus mimicking the need to represent unobservable content. A series of experiments with children aged between 3- and 5-years old

replicated the developmental divide between younger and older children, suggesting here that ToM is evident at some point around a child's fourth birthday. Similarly, those children that struggled with the false belief task also found the false photograph task difficult. On the basis of this result, Zaitchik (1990) goes on to propose that "mental representations may be hard not because they're *mental*, but because they are *representations*" (p. 61). This is couched in terms of difficulty in assigning conflicting truth states to a single referent: the representational content presents as one reflection of the world, yet this is at odds with the child's perception of the state of the world. Children may see photographs and beliefs as necessarily true; however, failure in these tasks may be because the veridicality of the representational sources is overridden by the child's own perception of the true state of affairs. This is something that is explored extensively in Chapter 5, where I argue that representing mental state content is cognitively different from representing non-mental state content.

Emergence, Expression and Executive Functions

As already alluded to, ToM follows a fairly predictable developmental path. In their influential meta-analysis of false belief tasks, Wellman et al. (2001) demonstrated that false belief reasoning shifts from below- to above-chance performance during the pre-school years, at around the age of four. This provides strong evidence that the nature of children's responses switches from incorrectly, but systematically, maintaining their own viewpoint as a point of reference, to being able to actively realise the competing, but contextually relevant, perspective of an agent. Wellman et al. (2001) outline two major theoretical interpretations in respect of this trajectory. Conceptual change accounts suggest that ToM concepts must first be acquired before they can be expressed. ToM is, therefore, characterised by dichotomy – it is a conceptual problem to be overcome. Proficiency with executive functions, for example,

coincides with the development of ToM but is not directly related to ToM ability. As a result, younger children may fail false belief tasks as they lack the necessary executive skills to allow the ToM concept to emerge. An alternative account argues, instead, for early performance. This proposal suggests that the expression of ToM is mediated by supporting processes, such as executive functions, and that variation in ToM performance across the lifespan reflects cognitive aptitude or limits with such processes (also see Apperly, 2012). Wellman et al. (2001) identified that children's performance in false belief reasoning could be enhanced by changes to the task protocol, such as making the agent's mental state more salient, or reducing the salience of the participant's own, real-world, perspective. Note, however, that performance was unaffected by variations in the referent on whom the representational content was assigned, such as puppets, dolls and real people. The characteristics of those performance-relevant task changes hint, not only, towards the presence of underlying control mechanisms in facilitating ToM, but also to the cognitive methods with which we approach mentalizing. Nevertheless, variation in task performance, through paradigmatic changes such as salience manipulations, could not sufficiently facilitate performance to shift children from being unsuccessful to successful in false belief reasoning; it simply served to account for noise at varying points over early childhood, outside of a relatively stable point of ToM acquisition. Thus, Wellman et al. (2001) argued for a developmental account of ToM, where development should be taken to refer to the expression, as it interacts with the concept, of ToM. In Chapters 2, 4 and 5, this theory is examined. Using data from adult participants, these chapters demonstrate that the difficulties experienced by children persist in adulthood. This suggests that executive function may, therefore, not simply be requisite to facilitate the emergence of ToM, but that executive mechanisms work alongside ToM on an ongoing basis throughout the lifespan.

The relationship between ToM and executive function has been studied extensively, with the majority of studies drawing a parallel between ToM and the maturation of control processes, particularly inhibitory control (see reviews by Carlson & Moses, 2001; Carlson, Moses, & Breton, 2002; Hughes, 1998; Perner & Lang, 1999). Leslie and Polizzi (1998) outline a model of false belief reasoning which attributes successful performance to a 'target shifting' mechanism. Working on the premise that a concept of belief is only meaningful if it typically reflects the truth, Leslie and Polizzi (1998) suggest that reasoning about a false belief requires representation of a non-default state which, therefore, attracts an inhibitory cost. On this basis, if inhibition fails, the default, true belief content is allowed and the agent is incorrectly assigned a reality congruent belief state. False belief reasoning, thus, takes on a sequence where the default true belief content is identified, and inhibition then follows to disengage and redirect attentional resources towards the alternative, false belief content. This model is defined in more general terms as enabling a shift from one salient, to a second less salient, target. This enabled Leslie and Polizzi (1998) to extend their theory to explain how a more complex ToM reasoning might successfully be applied to an agent, such as in the case of belief-desire reasoning. If we return to Maxi, one commonly overlooked feature of this task is that it is implicit that Maxi wants to find his chocolate. If, however, Maxi did not want to find his chocolate – perhaps he was hoping to avoid temptation from over indulging – the participant must take account of more than just Maxi's belief as to where he left his chocolate. Now, whether Maxi wants to locate his chocolate, too, becomes relevant. Here, the model predicts that one would first identify the target which is thought to contain the chocolate, so as to then go on to pinpoint the desired, chocolate-free cupboard. The participant must therefore disengage their attention in order to shift it from one target to another. On the basis that Maxi also held a false belief about the true location of chocolate, however, Leslie and Polizzi

(1998) propose that an effortful double inhibition must occur, in which the attentional shift from belief and desire content interact to cancel one another out.

The task analysis of inhibition in belief-desire reasoning proposed by Leslie and Polizzi (1998) provides an outline of how executive control might be exerted in ToM. In turn, this makes a strong prediction as to what specific ToM states would look like behaviourally. It also suggests that belief and desire would recruit a single neural mechanism to perform this task (Hartwright, Apperly, & Hansen, 2012). Overall, their model suggests that, compared with true belief and positive (approach) desire reasoning, false belief and negative (avoidance) desire reasoning will prove more effortful. This effect has been demonstrated in young children (Cassidy, 1998; Leslie & Polizzi, 1998) and adult participants (Apperly, Back, Samson, & France, 2008; German & Hehman, 2006; Hartwright et al., 2012). The fact that adults experience similar difficulty to young children is particularly interesting, as it suggests that a pure conceptual account of ToM is insufficient to explain difficulty with reasoning about beliefs. Although it is not unreasonable to assume that young children fail ToM tasks due to an underdeveloped concept of belief, it is unfathomable to apply this logic to an adult. When considered in the context of the evidence provided, such difficulty is likely, therefore, to reflect processes which are not purely incidental to task design. This is a core element that runs through the present thesis; what are the neurocognitive processes involved when representing these effortful ToM states? Chapter 2 begins to unpack this.

Difficulty with false belief reasoning may also, or instead, reflect interference from one's own knowledge of reality (Birch & Bloom, 2004; Birch & Bloom, 2007; Ruby & Decety, 2003). This theory stems from young children's propensity to systematically adopt their own viewpoint. For example, if shown that a tube of Smarties actually contains pencils, children of 3-years old describe the contents as if they had always known that there were

pencils inside, and as if others would also expect there to be pencils inside (Gopnik & Astington, 1988). Adults also tend to overestimate the extent to which others share their visual perspective (Keysar, Lin, & Barr, 2003; McCleery, Surtees, Graham, Richards, & Apperly, 2011), own general knowledge (Thomas & Jacoby, 2013), or outcome knowledge (Fischhoff, 2003). As previously outlined, pre-school aged children consistently and systematically adopt their own, privileged knowledge point when attempting to reconcile a false belief (Wellman et al., 2001). Similarly, adults demonstrate processing costs when imputing belief states that are distinct from their own conceptual perspective (Apperly et al., 2008; German & Hehman, 2006; Hartwright et al., 2012). Performance can be enhanced in ToM tasks where the salience of own perspective is reduced (Samson, Apperly, Kathirgamanathan, & Humphreys, 2005), or the salience of the agent's perspective is increased (Wellman et al., 2001). Collectively, these data suggest that our own perspective may interfere with our ability to assume the viewpoint of an agent (Birch & Bloom, 2004; Birch & Bloom, 2007; Ruby & Decety, 2003; Samson et al., 2005). In order to solve a classic unexpected transfer task, the participant has to suppress their own prepotent but, in the case of the task, incorrect true belief, whilst simultaneously holding action relevant information in working memory. If we return to Maxi, then, the difficulty may lie in knowing that Maxi's mother moved the chocolate to cupboard Y; our own, salient self perspective is a possible source of interference. Within this framework, in order to predict where Maxi will look for his chocolate, self knowledge of the real location must be resisted in order to select the empty, and therefore somewhat erroneous, location X. This task analysis explains that the behavioural difficulty that is evidenced with false belief reasoning, in particular, may be attributable to a "curse of knowledge" (Birch & Bloom, 2004; Birch & Bloom, 2007) or "reality bias" (Mitchell & Lacohee, 1991). In the case of young children, failure on false

belief tasks could reflect immature executive control systems, not simply an ignorance of belief states, or this extra effort may serve to exaggerate existing conceptual limitations (Birch & Bloom, 2007).

The bias towards the content of one's own perspective makes clear predictions about which ToM states will be more effortful for children and adults to impute. Specifically, it suggests that conflict will exist when there is incongruence between self and other perspectives, as is the case in classic false belief reasoning: we know that Maxi's chocolate is in cupboard Y, yet Maxi thinks his chocolate is in cupboard X. If Maxi, however, held a true belief – he saw his mum move the chocolate to cupboard Y – there is no difference between our views on the state of the world (with regards to the location of his chocolate). Note, however, that his desire to find or avoid his chocolate conflicts in no way with our own perspective. The fact that he wants to find his chocolate, or not as the case may be, is in no way logically incongruent with our viewpoint of the world as it really is (Hartwright et al., 2012, [Chapter 2 here]). The curse of knowledge, therefore, makes a prediction regarding epistemic mental states (Samson et al., 2005), where processing a false versus true belief, but not an avoidance versus an approach desire, will have a demonstrable neurocognitive cost. This is examined in detail in Chapters 2, 4 and 5.

A Social Cognitive Neuroscience of Theory of Mind

A breadth of research tools including fMRI, lesion data, comparative studies and neurophysiological techniques attempt to describe how ToM may be represented neurally. As a result, researchers have used an array of paradigms including mentalizing vignettes, static cartoons, interactive games, videos and animations to attempt to engage and modulate 'the social brain' (Carrington & Bailey, 2009). Earlier in this chapter, I loosely defined ToM as the

ability to know what it is in another person's mind. Within neuroscience, however, the data on which our perspective of an agent is based, in terms of physical or mental cues, has different neurocognitive consequences. Thus, a distinction between following bodily cues, sometimes termed 'motor ToM' (Agnew, Bhakoo, & Puri, 2007), and following mental cues, 'mentalizing', should be made (Van Overwalle, 2011). This is fundamental as two separate neural networks have been identified to subserve action observation versus mentalizing. Not only do these comprise different cortical regions, but the properties of the neurons within these regions are quite different. This is of consequence for theorising about how the respective element of social cognition is achieved. Though not the focus of the current thesis, physical action is rarely separate from the social context (Hamilton, 2013a), and observations of such actions do not float free from ToM (Gallese, 2007; Gallese & Goldman, 1998). Consequently, I now provide a brief overview of the neural system involved in 'motor ToM', the Mirror Neuron System (MNS). The MNS has been typically studied during the execution and observation of movement, for example, the participant might be required to lift their index finger as if to tap the table, then watch an agent making the same movement. The MNS comprises neurons which reflect a unique 'mirroring' property; thus, these neurons discharge both during the execution and observation of an action (Iacoboni & Dapretto, 2006; Rizzolatti & Craighero, 2004). The property of mirror neurons provides a motor theory for social development (Vanderwert, Fox, & Ferrari, 2013), for example, in terms of explaining early infant imitative action as a precursor to more complex social cognition (Gallese & Goldman, 1998; Vanderwert et al., 2013). Mirror neurons were initially identified using single unit recording in macaques (Casile, 2013); however, neurophysiological evidence from electroencephalogram (EEG) and TMS studies suggested that regions of the human premotor cortex (PMC) reflect similar mirroring properties (Rizzolatti & Craighero, 2004), for

example, by using the finger tap approach described previously. Brain imaging has corroborated these data, but additionally suggests that inferior frontal gyrus pars opercularis (IFG_{op}), the inferior parietal lobule (IPL) and superior temporal sulcus (STS) may also demonstrate this action-observation effect in human subjects (Iacoboni & Dapretto, 2006; Rizzolatti & Craighero, 2004; Vanderwert et al., 2013).

In terms of how the MNS and the properties of mirror neurons may relate to mentalizing, researchers have speculated that a deficit in the MNS might explain ToM dysfunction in autism. For instance, if the MNS were disrupted, this would preclude the ability to generate internal representations of others through embodied simulation; therefore, preventing any experiential understanding of others (Iacoboni & Dapretto, 2006). A recent review, however, found weak evidence for this (Hamilton, 2013b). Nevertheless, the MNS is responsive to more than biological movement. For example, the MNS is significantly more active when viewing hand actions within, versus without, a social context, such as grasping the handle of a cup as if to drink. This suggests that the MNS is automatically responsive to social intention, not just the act of grasping (Iacoboni et al., 2005). Furthermore, co-activation of the MNS alongside core ToM brain regions suggests that the MNS might co-opt ToM regions when identifying social intention through action kinematics (Becchio et al., 2012). When an agent performs an unusual or implausible physical act, however, such as switching on a light using their knee, ToM brain regions are recruited in place of the MNS. Thus, the MNS appears to generate and operate on the basis of movement-related schema, where readily interpretable physical acts are assimilated in order to prepare a socially relevant response. Where a physical behaviour does not return a match to the movement schema, higher-order ToM brain regions are brought in to facilitate social understanding.

Whilst the MNS enables recognition of a goal through perceived physical action, the ToM network utilises 'social intelligence', or mental state content to make sense of an agent's behaviour (Van Overwalle & Baetens, 2009). Neural regions most widely agreed to constitute a core mentalizing network are the TPJ and mPFC (Bzdok et al., 2012; Carrington & Bailey, 2009; Gallagher & Frith, 2003; Saxe & Kanwisher, 2003; Van Overwalle, 2009); however, ToM also regularly recruits the temporal poles, precuneus and lateral prefrontal regions (Lieberman, 2007; Mar, 2011; Spreng et al., 2009). Though there is relative consensus that TPJ and mPFC constitute 'core' ToM regions, less is understood about how they respond to different ToM contexts, or how other neural regions, particularly those which support executive processes, might interact with TPJ and mPFC. Nevertheless, from the developmental literature reviewed earlier, it should be clear that ToM is likely to engage executive functions. The preceding text highlighted that the pre-school years reflect an important transitional period in the development of ToM where, at around the age of 4-years, most children demonstrate an understanding of other people's mental states. It is unlikely to be a coincidence that significant cortical reorganisation occurs during the pre-school years. This consists of both an increase in cortical surface area within the prefrontal and temporal association cortices, alongside a reduction in cortical thickness in medial prefrontal regions (see Brown & Jernigan, 2012 for a review). Nonetheless, as will be argued throughout this thesis, both lateral and midline frontal regions play an important role in supporting an adult ToM.

A common approach to studying the neural basis of ToM follows the developmental approach. Accordingly, the process of representing mental content, such as an agent who holds a false belief, versus representing structurally matched, non-mental content, like a false photograph, is examined. Saxe and Kanwisher (2003) formulated this approach into a 'ToM

localizer' task. Below are two example stimuli from this localizer. The participant reads a short story which is followed by a question requiring a true or false response. Participants complete a series of these false belief and false photograph vignettes whilst fMRI data are collected. The localizer task works on the premise that the two types of vignette differ only in the type of representation that is required. Thus, those brain regions which are preferentially responsive to false belief, over and above false photograph reasoning, are thought to reflect differential processing between representing mental, versus non-mental, content. For example,

False belief format

Anne made lasagne in the blue dish. After Anne left, Ian came home and ate the lasagne. Then he filled the blue dish with spaghetti and replaced it in the fridge.

–

Anne thinks the blue dish contains spaghetti.

True *False*

False photograph format

A photograph was taken of an apple hanging on a tree branch. The picture was not viewed for half an hour. In the meantime, a strong wind blew the apple to the ground.

–

The developed photograph shows the apple on the ground.

True *False*

Vignettes from **Saxe and Andrews-Hanna (n.d.)**

Repeated use of this localizer demonstrates consistent recruitment of TPJ, mPFC, precuneus and temporal poles (e.g., Aichhorn et al., 2009; Hartwright et al., 2012; Mitchell, 2007; Perner, Aichhorn, Kronbichler, Staffen, & Ladurner, 2006; Saxe & Powell, 2006; Saxe & Wexler, 2005; Scholz, Triantafyllou, Whitfield-Gabrieli, Brown, & Saxe, 2009; Young, Dodell-Feder, & Saxe, 2010). Therefore, this task has been used extensively to interrogate the functional profiles of these regions. This approach has highlighted that, compared with the other regions identified by the localizer, only right TPJ appears to be selectively responsive to representing mental state content, such as the beliefs and desires of an agent, and that it is not simply responsive to the mere presence of social content or human actors (Saxe & Kanwisher, 2003; Saxe & Wexler, 2005). As the localizer task identifies parts of the brain that seem to be specifically engaged for ToM, this tool is used alongside some of the novel experiments presented in this thesis.

Neural Regions of Interest

Likely due to the breadth of disciplines that utilise neuroimaging methods, there is no single, standardised neuroanatomical labelling system (Devlin & Poldrack, 2007). Limited spatial specificity associated with tools such as EEG or Positron Emission Tomography (PET) necessitates labelling on a macroanatomical level, where regions may be described using lobular terms, or in respect of gross sulcal or gyral anatomy. High resolution methods including functional and structural MRI, on the other hand, attract labelling of functional or connectional architecture. Due to the inconsistent neuroanatomy nomenclature, it is challenging to qualitatively evaluate commonalities across studies localizing ToM or ToM processes. The experiments outlined within this thesis focus on neural regions which are regularly implicated in ToM, the TPJ and mPFC, alongside those which are more typically

associated with executive control, the vlPFC and dmPFC. I make special effort to delineate these relatively large anatomical regions, in terms of both defining where in vlPFC might be most relevant for ToM, and demonstrating how a differing ToM context yields a functional divide in mPFC. For clarity of discussion, Fig. 1. highlights anatomical location and labelling of the core neural regions which are evaluated within this thesis.

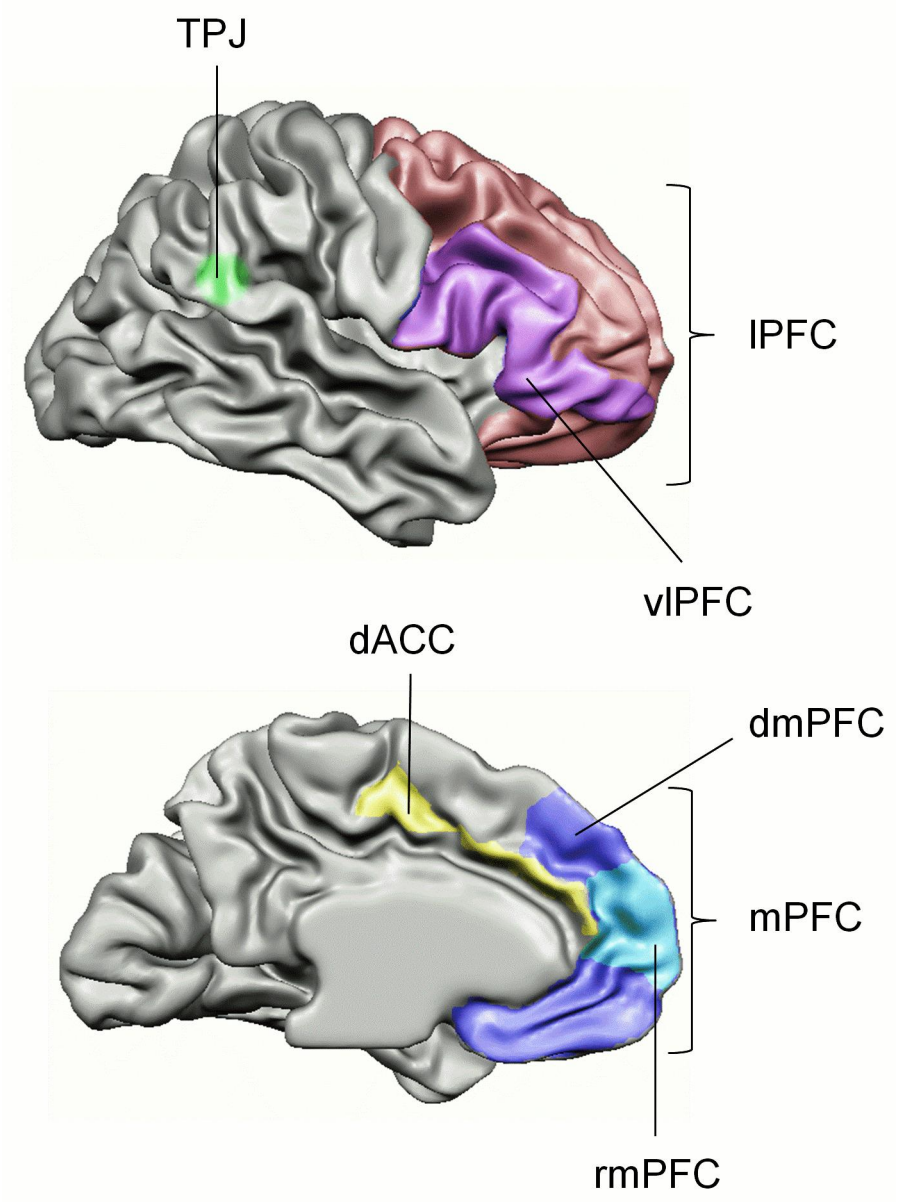

Fig. 1. Neural Regions of Particular Relevance
Coloured regions reflect brain areas which receive regular discussion throughout this thesis. TPJ = temporoparietal junction; PFC = prefrontal cortex; l = lateral; v = ventral; m = medial; r = rostral. dACC = dorsal anterior cingulate cortex

Thesis Structure

The data presented within this thesis detail how specific mental states are represented in ToM and executive control regions. Novel task analyses are presented from which specific hypotheses regarding how experimental manipulation will modulate core brain regions are tested. These data suggest that ToM processing is affected by differences in executive demands that are associated with certain ToM contexts and concepts. My work highlights that existing descriptions of the neurocognitive bases of specific belief and desire states require modification. Such modifications should include reference to how changes within mental state concepts, and the wider representational context, affect the distribution and functional profile of resources used within the brain.

Each empirical chapter comprises a self-contained manuscript. Each chapter, therefore, typically comprises an overview of any core terminology and background literature, a detailed account of the methods used and results obtained, plus a discussion in the context of the particular question(s) of interest. The first empirical chapter, Chapter 2, evaluates two theoretical frameworks concerning the role of executive control in ToM. The first framework suggests that the ability to represent the perspective of self, versus that of another, necessarily requires executive control. This process is postulated to be more effortful when the thoughts and beliefs of an agent are incompatible with our own. The second suggests that executive selection resources are required more generally, in attentionally demanding situations, for example, when circumstances elicit competing information streams. By using a novel, belief-desire reasoning task alongside an extensively published ToM localizer task, a neural basis for the processes underlying each of these frameworks is proposed. In particular, a role for vlPFC in the inhibition of self perspective, and dmPFC, particularly ACC, in supporting more general control demands, is outlined. The response of TPJ to this manipulation is also described. This is explored

in the context of what these results suggest about the function of TPJ in the wider ToM sphere. The overall results are then contextualised within how children and adults might apply a ToM.

Intrigued by the absence of neural activation in mPFC in Chapter 2, a region often thought critical for mentalizing, the third chapter presents a task analysis of ToM. Here, three processes – representation, control and reasoning – are explored. These processes are suggested to feature in ToM, in varying degrees, according to the representational context. The ToM task outlined in Chapter 3 sought to explain the absence of mPFC in terms of this proposed contextual effect. A new task is outlined, which adopted a minimal change to the paradigm that features in Chapter 2. This enabled close inspection of how variation in the mode of reasoning modulates core ToM brain regions, and was used to demonstrate a functional dissociation in mPFC. Here, the theory that two distinct types of reasoning can be employed when making a representational judgment is described.

The experiment presented in Chapter 2 postulates that vlPFC supports inhibition of self perspective in ToM. Chapters 4 and 5 focus in on this region, in order to further explore its functional profile. The study described in Chapter 4 used TMS. Perturbation techniques such as TMS enable a causal inference regarding the role of the stimulated region, by inducing a 'virtual lesion' and any associated behavioural consequences (Pascual-Leone, Walsh, & Rothwell, 2000). Thus, TMS permits more powerful conclusions regarding the role of vlPFC in ToM. As the experimental protocol differs from the other methods presented in this thesis, the merits of TMS are briefly reviewed in relation to neuropsychological and fMRI evidence. Consideration is given to the application of TMS to vlPFC, in particular. The effects of TMS were shown to be highly variable across subjects. The failure to elicit the hypothesised effects is reviewed in relation to physiological issues – such as difficulties with

localization and neural efficacy – and paradigmatic concerns, which have arisen following recent evidence which now brings into question the TMS protocol that was used.

The final empirical chapter uses the same ToM localizer task that features in Chapter 2. However, in order to clarify the role of vlPFC in ToM, Chapter 5 describes a modification to the original localizer that systematically varies the salience of a participant's own perspective, when making a representational judgement. In doing so, modulation of vlPFC was demonstrated. The study outlined in Chapter 5 identified a functional dissociation in vlPFC, where salience of self perspective was only relevant when representing mental versus non-mental content. In light of this, Chapter 5 describes existing theories regarding possible processing differences between mental and non-mental representation. These theories are integrated into a unified proposition which tentatively aims to explain the role of vlPFC in ToM and non-mental representation.

The closing chapter of this thesis provides a summary of the major findings. In Chapter 6, an attempt is made to integrate these results into a flexible framework. Such a framework reflects how the data presented in this thesis might extend our interpretation of what constitutes the core ToM network, in terms of the neurocognitive bases of context specific ToM processes.

CHAPTER 2:

MULTIPLE ROLES FOR EXECUTIVE CONTROL IN BELIEF-DESIRE REASONING: DISTINCT NEURAL NETWORKS ARE RECRUITED FOR SELF PERSPECTIVE INHIBITION AND COMPLEXITY OF REASONING[1]

[1] This chapter is published: Hartwright, Charlotte E., Apperly, Ian A., & Hansen, Peter C. (2012). Multiple roles for executive control in belief-desire reasoning: Distinct neural networks are recruited for self perspective inhibition and complexity of reasoning. *NeuroImage, 61*(4), 921-930. doi: 10.1016/j.neuroimage.2012.03.012

ABSTRACT

Belief-desire reasoning is a core component of 'Theory of Mind' (ToM), which can be used to explain and predict the behaviour of agents. Neuroimaging studies reliably identify a network of brain regions comprising a 'standard' network for ToM, including temporoparietal junction and medial prefrontal cortex. Whilst considerable experimental evidence suggests that executive control (EC) may support a functioning ToM, co-ordination of neural systems for ToM and EC is poorly understood. We report here the use of a novel task in which psychologically relevant ToM parameters (true versus false belief; approach versus avoidance desire) were manipulated orthogonally. The valence of these parameters not only modulated brain activity in the 'standard' ToM network but also in EC regions. Varying the valence of both beliefs and desires recruits anterior cingulate cortex, suggesting a shared inhibitory component associated with negatively valenced mental state concepts. Varying the valence of beliefs additionally draws on ventrolateral prefrontal cortex, reflecting the need to inhibit self perspective. These data provide the first evidence that separate functional and neural systems for EC may be recruited in the service of different aspects of ToM.

INTRODUCTION

The capacity to reason about the mental causes of action, termed 'mentalizing' or exercising a 'Theory of Mind' (ToM), has received considerable interest from social neuroscientists over the last decade. Much attention has been given to identifying which, if any, brain regions should be considered as specialised for ToM. This work has made considerable progress in identifying possible contenders, and converges on the importance of a network of brain regions including temporoparietal junction (TPJ) and medial prefrontal

cortex (mPFC) (Carrington & Bailey, 2009; Lieberman, 2007; Mar, 2011; Van Overwalle, 2009). TPJ has been identified in the majority of neuroimaging studies of ToM, and appears selectively responsive when representing mental states such as beliefs, desires and intentions, over and above representation of physical states, personality traits or dispositions of the person, and above non-mental representations, such as photographs (Aichhorn et al., 2009; Saxe & Kanwisher, 2003; Saxe & Wexler, 2005). mPFC is also identified in most neuroimaging studies of ToM, though its activity may be less specific to mental state representation (Amodio & Frith, 2006), and may be most strongly recruited when reflecting on more enduring mental states, such as personality traits and social or moral beliefs (Van Overwalle, 2009), or when making inferences under conditions of high uncertainty (Jenkins & Mitchell, 2009). The strong convergence of neuroimaging data has lead to a general consensus that TPJ and mPFC constitute the 'core' network for ToM, and that the functions they support are the most psychologically important for understanding ToM.

ToM has been studied most extensively using false belief tasks. A classic paradigm, the object transfer task, requires participants to make a prediction about the behaviour of a character, based upon the character's belief and desire at that point in time. A typical experimental sequence outlines a protagonist putting an object into location A. They then leave the scene. Whilst the protagonist is away, the object is transferred to location B. The character then returns, wishing to find the object but holding a false belief about its location (Wimmer & Perner, 1983). In order to successfully identify where the protagonist will look for the object, it is necessary for participants to infer the character's false belief about the object's location and predict the character's action on the basis of their false belief, while resisting interference from their own privileged knowledge of the object's true location, and what the right course of action would be. This task analysis leads to the expectation that

successful ToM will not only require processes that might be specific to inferring and representing the mental states of others, but also processes for executive control (EC) to ensure that the correct information is selected for inferring mental states and predicting actions. It follows, then, that a complete account of the neural basis of ToM must also include brain regions associated with these sorts of control processes. To date, however, the brain bases of EC in ToM have been little explored.

Numerous researchers have noted that executive ability appears to contribute significantly to proficiency with ToM (Carlson & Moses, 2001; Carlson et al., 2002; Carlson, Moses, & Hix, 1998; Friedman & Leslie, 2004; German & Hehman, 2006; Leslie, German, & Polizzi, 2005; Leslie & Polizzi, 1998; Perner & Lang, 1999; Wellman et al., 2001). For example, in the classic false belief paradigm mentioned above, children under the age of four seem unable to overcome their own knowledge of where the object really is. As a result, they consistently state that the protagonist will look for the object in the object's true location. Younger children, however, may sometimes pass the false belief task if the true location of the object is made less salient (Carlson et al., 1998; Wellman et al., 2001). These kind of 'egocentric errors', sometimes referred to as the 'curse of knowledge' (Birch & Bloom, 2004; Birch & Bloom, 2007) or a 'reality bias' (Mitchell & Lacohee, 1991), have also been observed in older children and healthy adults (Bernstein, Atance, Loftus, & Meltzoff, 2004; Birch & Bloom, 2007) and appear to reflect the need to exert EC to solve such tasks (see e.g., Apperly, Samson, & Humphreys, 2005; Apperly, Samson, & Humphreys, 2009 for relevant discussions).

Two broad theoretical frameworks have been proposed concerning the role of EC in ToM. The first suggests that EC is necessary when a perspective difference between self and other exists, as is the case of false belief or conflicting desires. For example, knowledge of the

true location may interfere with the ability to select the believed, or false, location. As a result, the self perspective must be inhibited in order to assume the perspective of the other (Ruby & Decety, 2003; Samson et al., 2005). This theory arises from behavioural observations of young children's propensity towards responding with their own knowledge, and data suggesting that performance in ToM tasks can be manipulated by varying the salience of self perspective (Carlson et al., 1998; Wellman et al., 2001). A growing literature suggests that the ventrolateral prefrontal cortex (vlPFC) may support this process of 'self perspective inhibition'. For example, Vogeley et al. (2001) identified that the right inferior frontal cortex, particularly right inferior frontal gyrus (rIFG), was modulated by varying the importance of self in a fictional scenario. This finding was later supported by a case study which demonstrated that damage to right vlPFC, including rIFG, resulted in interference from self perspective when attributing beliefs to others. In this particular case, the patient was able to solve ToM tasks where his own perspective was less salient, but failed ToM tasks where a clear incongruence between self and other knowledge state existed (Samson et al., 2005). Using false belief tasks from Samson et al.'s study and a stop-signal test of EC, a further study showed that the same ventral region of IFG was recruited bilaterally in healthy adults for both general response inhibition, and when contrasting false belief tasks that made high versus low demands on the inhibition of self-perspective (van der Meer, Groenewold, Nolen, Pijnenborg, & Aleman, 2011). Finally, a recent study of visual perspective-taking showed an ERP component over right fronto-lateral cortex that was sensitive to differences between self and other perspectives (McCleery et al., 2011). These studies provide converging evidence that a functioning ToM is supported by regions outside of the 'standard' ToM network, and that the inferior frontal cortex – particularly vlPFC – may be an important, but overlooked,

region involved in inhibition of self perspective. Notably, however, none of these studies examine the role of EC in reasoning about conative mental states, such as desires.

The second theory of the role of EC in ToM, proposed by Leslie and colleagues (Friedman & Leslie, 2004; Friedman & Leslie, 2005; Leslie et al., 2005; Leslie & Polizzi, 1998), extends beyond belief attribution to include the varying demands of desire reasoning. It is implicit in the standard false belief task that the character wishes to locate the object. However, if the agent holds a desire to avoid the object, both children and adults suffer further difficulty in false belief tasks (Apperly, Warren, Andrews, Grant, & Todd, 2011; Cassidy, 1998; German & Hehman, 2006). Moreover, like false belief reasoning, proficiency with avoidance desire coincides with the development of executive abilities. Leslie and colleagues explain this in terms of a shared inhibitory component for negatively valenced[2] mental states, such as false belief and avoidance desire. They suggest that, for both false belief and avoidance desire reasoning, participants are required to select from competing responses and inhibit the prepotent response (e.g., true versus believed location / desired versus undesired location). Consequently, false belief and avoidance desire states may draw on a domain-general 'selection processor', in order to direct executive selection resources in attentionally demanding situations. Importantly, avoidance desire (i.e. "desire to avoid") can concern objects or situations that are either intrinsically desirable or undesirable from the participant's own point of view. Hence, variation in this valence of desire does not reduce to a question of whether the participant shares the character's desire: indeed desire valence and self-other congruence of desire are logically orthogonal factors. To explain these findings, Leslie and colleagues suggest that EC has a more general role in ToM that is not restricted only to cases

[2] These variations in belief and desire both vary the difficulty of the belief-desire reasoning task. However, beliefs in the current study varied in terms of their consistency with the participant's self-perspective (true beliefs versus false beliefs), whereas desires varied only in terms of whether the target character liked or disliked the food. Therefore we use the term "valence" to refer collectively to these variations, so that true beliefs and desires for foods are described as "positively valenced" and false beliefs and desires to avoid foods are described as "negatively valenced".

that require inhibition of self-perspective. Previous neuroimaging studies examining EC more generally in ToM have typically used separate tasks to identify ToM and EC regions. These indicate some overlap between neural regions recruited for EC tasks and false belief reasoning, extending beyond IFG to include anterior cingulate cortex (ACC), frontal operculum (FO) and frontal eyefields (FEF) (Rothmayr et al., 2011; Saxe, Schulz, & Jiang, 2006; van der Meer et al., 2011).

Whilst interest in the neural basis of executive function in ToM is growing, most previous studies are limited in their ability to cast light on the role of executive function in ToM. Most have sought to identify neural regions involved *only* in ToM by comparing activation observed in a ToM condition with that in a non-ToM control condition. Such approaches may enable powerful tests of hypotheses about brain regions that are domain-specific for ToM, but run the risk of subtracting out activation that is critical for understanding how ToM is achieved in the brain. A fruitful alternative approach is to manipulate psychologically relevant factors within a ToM task (for a discussion of these issues see Friston & Henson, 2006; Saxe, Brett, & Kanwisher, 2006). Surprisingly few previous studies of ToM, however, have attempted such manipulations. Sommer et al. (2007) provide one of the few direct comparisons between true and false belief reasoning. In their nonverbal task, participants viewed a series of cartoons which depicted a true or false belief scenario analogous to the object transfer task outlined earlier. Regions which were more responsive to false belief over true belief attribution included the right TPJ, ACC and right lPFC. The reverse contrast only identified the superior frontal gyrus, which is in contention with the view that TPJ is an essential component when attributing any transient mental state (seeVan Overwalle, 2009). These data indicate that false belief reasoning might recruit EC regions. However, they are difficult to interpret with confidence, because is not clear whether

participants were solving the contrasting true belief condition by mental state ascription or by simply referring to the true state of affairs (Aichhorn et al., 2009). Consequently, further examination of these two mental states is warranted, where attending to a protagonist's mental state is made unavoidable in both true and false belief reasoning. This was the case in the current study.

The neural basis of conative states such as desires has been studied less extensively. Hooker, Verosky, Germine, Knight, and D'Esposito (2008) examined neural activation when making empathic judgements for characters with varying perspectives. More directly relevant to the current study, Abraham, Rakoczy, Werning, von Cramon, and Schubotz (2010) had participants read a series of short vignettes which varied the valence of belief and desire: either an agent's belief turned out to be true or false, or an agent's desire turned out to be fulfilled or unfulfilled. The vignettes were followed by a yes/no question in which participants judged how the agent would feel about the true state of affairs. Their results were broadly consistent with the existing literature and showed recruitment of key mentalizing areas including TPJ and mPFC for both the belief and desire conditions compared to a non-ToM reasoning task. An analysis of the overall effect of valence (of both belief and desire) identified activation in mid-line structures, including mPFC and posterior cingulate cortex. This study is interesting because it attempts to separate the demands of belief and desire reasoning into different experimental conditions. However, this also leads to limitations. Firstly, it is unclear whether this separation can be entirely successful, since judging an agent's feelings on the basis of his belief may lead participants to think about his desire even though they were not asked to. Likewise, judging an agent's feelings on the basis of her desire may lead participants to think about her belief. Secondly, it is unclear how such conditions relate to the canonical forms of ToM reasoning, in which we combine information about both

belief and desire to predict or explain an agent's action. For this reason, the current study followed the longstanding literature on ToM in children by asking participants to predict a character's actions on the basis of his belief and desire.

We deployed a novel task (Apperly et al., 2011) based upon the object transfer action prediction ToM task, from which there are already considerable behavioural data (e.g., Friedman & Leslie, 2005; Wellman et al., 2001; Wimmer & Perner, 1983). Our previous work has shown this design to be able to detect differences in reaction time and error rate when participants predicted an agent's action on the basis of true versus false beliefs and a desire to approach versus avoid an object. This task allowed us to look specifically at neural activation during the decision making phase during which these behavioural effects are observed. The novel task comprises an orthogonal design whereby belief (true/false) and desire (approach/avoid) states are manipulated within a single, within-subjects experiment. The use of this factorial design enabled a whole brain analysis to isolate any neural regions that were modulated either by the valence of belief state, or by the valence of the desire state, or both. In doing so, the present study sought to address three key questions:

Do our factors of Belief-Valence and Desire-Valence recruit any regions of the ToM network?

It is entirely possible that our factors of Belief- and Desire-Valence would not recruit any regions of the ToM network, because beliefs and desires feature in all of our experimental conditions. It is important to emphasize that the present task and analyses were not designed to identify regions that are specifically involved in representing beliefs or desires in comparison with non-ToM reasoning, but instead were designed to identify those regions that are responsive to variation in the valence of either belief or desire during an action prediction.

This is informative because, as reviewed above, previous work shows that the valence of beliefs and desires makes a critical contribution to the difficulty of belief-desire reasoning for both children and adults.

If our factors of Belief-Valence and Desire-Valence recruit regions of the ToM network, is this just because those regions are involved in attention/executive control, not because they are involved in ToM per se?

Although the literature converges on identifying brain regions that are consistently associated with ToM, the role of these regions remains controversial. On one view, at least some regions – in particular, some regions of right TPJ – are activated during ToM tasks because they are specifically involved in ToM (e.g., Saxe & Kanwisher, 2003; Scholz et al., 2009). On another view, such activation merely reflects the allocation or reorientation of attention, which is known to be a function of TPJ, and is a confounding feature of many ToM tasks (Mitchell, 2007; Rothmayr et al., 2011). The need for care on this question is emphasised by a recent structural imaging study which demonstrated that TPJ can be subdivided in terms of its connectivity with other brain regions associated respectively with ToM and attention (Mars et al., 2012). To address this issue, we used a separate ToM localizer task (Saxe & Kanwisher, 2003) alongside our novel task. This localizer contrasts brain activation observed during false belief trials with that observed during closely-matched false photograph trials. It is widely agreed that false photograph tasks are an excellent match for most of the confounding demands that false belief tasks make on memory, EC and attention (e.g., Aichhorn et al., 2009; Saxe & Powell, 2006), so although interpretation of this localizer remains controversial (e.g., Mitchell, 2007; Young, Dodell-Feder, et al., 2010), it is currently the best method available for identifying brain regions that might be specifically

involved in ToM. By using the localizer alongside the belief-desire task, we were able to explore the neural signature of specific belief and desire states in those voxels within TPJ that appear to be specifically responsive to mental representation.

Do we observe differential activation of executive control regions due to the Belief-Valence factor compared with the Desire-Valence factor?

The two theories of EC in ToM reviewed earlier make alternate, but not incompatible, predictions about the pattern of brain activation in belief and desire reasoning. Firstly, Leslie and colleagues' executive performance account of ToM (Friedman & Leslie, 2004; Friedman & Leslie, 2005; Leslie et al., 2005; Leslie & Polizzi, 1998) posits that both avoidance desire and false belief reasoning recruit common executive resources for the selective control of attention. If this theory is correct, negatively valenced belief and desire states will draw on the same executive regions. Secondly, if EC is involved in self-perspective inhibition (McCleery et al., 2011; Samson et al., 2005; van der Meer et al., 2011), then we should expect to see different recruitment of brain areas for the factors of Belief-Valence and Desire-Valence. This is because false belief trials are thought to make higher demands on self-perspective inhibition than true belief trials, whereas there is no systematic variation in the need for self-perspective inhibition when the agent has a desire to avoid rather than approach the object.

METHOD

Participants

Twenty healthy adults participated in both of the fMRI experiments. All gave informed ethical consent and were given course credit or a small honorarium for their

participation. The study had appropriate research ethics approval from the University of Birmingham. One participant was excluded from all analyses due to poor behavioural task performance during scanning[3]. The remaining 19 participants were included in all analyses (6 male, 13 female; age range 18-39, $\overline{X}$ age = 25 years). All participants were strongly right handed, measured with a modified form of the Annett Handedness Questionnaire (1970), and were proficient English speakers.

Materials and Procedure

Pre-screen

Suitability to participate was determined several days prior to collecting the neuroimaging data. The Wide Range Achievement Test – Third Edition (WRAT-3) Reading Scale was administered to screen for reading disabilities and ensure reading proficiency commensurate with the experimental tasks. The participants then completed a computer based interactive training session which gave an overview of the belief-desire reasoning experiment. They then attempted one block of experimental trials outside of the MRI scanner. Only participants who performed above chance at $p < 0.05$ on this pre-test block took part in the fMRI experiments (see Appendix 1 for detailed participant screening information). Of twenty three prospective participants, 3 individuals (3 female; age range 19-24, $\overline{X}$ age = 23 years) were unable to perform the Belief-Desire Reasoning experiment to above chance at the pre-screen stage and thus did not participate in any of the fMRI experiments.

[3] Based on a binomial distribution at $p < 0.05$, this participant performed below chance on the belief desire task.

Belief-Desire Reasoning Experiment

The main experiment was based on a paradigm devised by Apperly et al. (2011) which was revised for use as an event-related design within the MRI scanner. Pilot work using the revised paradigm confirmed that experimental timings were appropriate in that the participants were able to perform the task to a high degree of accuracy (> 90% correct trials). The experiment utilised an orthogonal design which had four equally occurring conditions that were based on a protagonist's belief state (true (B+) or false (B-)) and desire state (approach (D+) or avoid (D-)). By varying the protagonist's beliefs and desires four conditions were created: B+D+, B+D-, B-D+ and B-D-. Note that immediately prior to participating in the main experiment all participants completed one further practice block outside of the MRI scanner so as to refamiliarise themselves with the main experimental task. None of the pre-test or practice trials were used in the main fMRI experiment.

The experiment required participants to predict which one of two different coloured boxes a character would open based on a scenario in which the character would seek out food they love and avoid food they hate (Fig. 2).

A male protagonist, introduced during the training and practice sessions as Simon, was always used with male participants, whereas a female character, Sally, was always used with female participants. Each scenario consisted of three centre justified statements followed by a picture response probe then rest. Statements were separated by a fixation period of 400 ms. A variable interstimulus interval was used (range = 9000-14000 ms, $\overline{X}$ = 11500 ms), during which a small fixation dot was displayed. The temporal order of the statement types was randomised, but all scenarios contained one belief statement (e.g., he thinks the chips are in the red box), one desire statement (e.g., he loves chips) and one reality statement (e.g., the chips are in the blue box). This design meant that participants were always explicitly told the

character's belief, whether the belief was true or false. Moreover, randomisation of the statement order ensured that participants needed to encode the character's true belief on at least the 50% of trials on which they did not already know the object's true location. In these ways our design addressed the weakness of earlier studies in which participants could safely ignore a character's beliefs on true belief trials, relying instead on their own knowledge of reality.

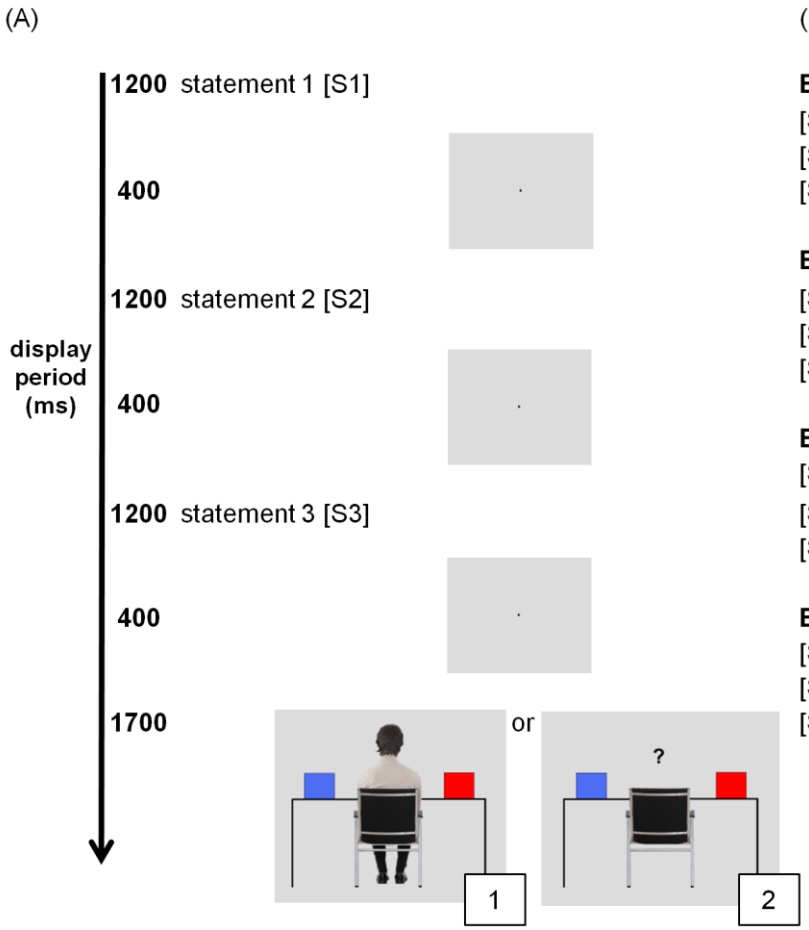

Fig. 2. Method: Belief-Desire Reasoning Paradigm
(Panel A) Experimental sequence of a single trial. Response probe 1 is an example of the image displayed for a trial of interest, picture 2 is an example of the response probe displayed during the anti-strategy trials. Note that the white numbered boxes were not part of the stimuli.
(Panel B) Example trial sequences for each of the four conditions. The order of statement types were randomised for each trial.

The statements were followed by a response probe. If the protagonist appeared in the response probe, participants indicated whether the character would open the left or the right box based on the agent's belief-desire state, using a two button box placed in their left hand[4]. These were the trials of interest and made up two thirds of the overall number of trials presented. In the other one third of trials, the protagonist was replaced with a question mark in the response probe. In this instance, participants responded by giving the true location of the food. These anti strategy trials were used to ensure that the participants had to attend to all three statements, and did not form any part of the analyses presented within the present paper. Twelve different food types were used, which were consistently "loved" or "hated" by the on screen protagonist. Food preferences were counterbalanced so that half of the participants saw one consistent set of preferences, whereas the other half saw the opposite preferences. The correct response corresponded to the left and right box an equal number of times. Participants completed four blocks of trials, each of which contained 24 trials (16 trials of interest, 8 anti strategy trials). Each block lasted 7 minutes 22 seconds which included an initial instruction and final thank you screen.

Theory of Mind Localizer Experiment

The localizer task was substantially based on the experimental procedure devised by Saxe and Kanwisher (2003). Stimuli consisted of a subgroup of the current localizer stories (see Saxe & Andrews-Hanna, n.d.), some of which were anglicised for use in the present experiment. Participants read a total of 24 short vignettes which referred to either a protagonist's false belief (FB) or an outdated physical representation, such as the false photograph scenario (FP). Each vignette was displayed for 10 s, which was followed for 4 s

[4] The left hand was used to differentiate potential language related neural activity from motor activity

with a short true or false question about the preceding story. This required participants to make a response using a two button box that was placed in their left hand. Stories alternated between FB and FP and were interleaved with a 13.5 s rest period. The localizer experiment comprised of four blocks of six trials, each containing three of each type of story. Participants were given four practice trials immediately prior to scanning to familiarise themselves with the localizer task.

Data Acquisition and Analysis

Neuroimaging Data Acquisition and Processing

Each participant's data were acquired during a single scanning session using a 3T Philips Achieva scanner. All stimuli were presented using Presentation software (Neurobehavioral Systems, CA) which also recorded the behavioural response data simultaneously. Participants completed two blocks of the main belief-desire experiment followed by all four blocks of the localizer task and the remaining two blocks of the main experiment. 177 T2*-weighted echo-planar imaging (EPI) volumes were obtained per block of the belief-desire experiment and 77 EPI volumes were acquired for each block of the localizer task. Both tasks utilised the same general imaging parameters to achieve whole brain coverage (TR = 2.5 s, TE = 35 ms, acquisition matrix = 96 x 96, flip angle = 83°, voxel size = $3x3x3mm^3$). EPI images consisted of 42 axial slices that were obtained consecutively in a bottom up sequence. High resolution T1-weighted structural images were acquired following collection of the functional data ($1x1x1mm^3$ isotropic voxels).

Preprocessing and statistical analyses of the data were performed using the FMRIB software library (FSL version v.5.98; FMRIB, Oxford, www.fmrib.ox.ac.uk/fsl). For both experiments, initial preprocessing of the functional data consisted of slice timing correction,

and motion correction using rigid body transformations (MCFLIRT). The blood oxygen level dependent (BOLD) signals were high-pass filtered using a Gaussian weighted filter of 30 s for the belief-desire task and 21 s for the localizer task. The BOLD data were then spatially smoothed using a 5mm full-width-half-maximum kernel. The functional data were registered to their respective structural images and transformed to a standard template based on the Montreal Neurological Institute (MNI) reference brain, using a 6-DoF linear transformation (FLIRT).

Belief-Desire Reasoning Experiment Analysis

The functional data resulting from the four conditions were modelled as four explanatory variables (EVs) of interest: B+D+, B+D-, B-D+, B-D-. To focus on the decision making phase of the sequence, the onset of each event was time locked to when the participant made a button response for each trial. Each EV comprised an arbitrary duration of 100 ms. The EVs were convolved with a gamma derived haemodynamic response function (HRF) within a general linear model framework (GLM). Motion parameters were treated as regressors of no interest in order to account for unwanted motion effects. The sentence phase was modelled as a regressor of no interest and orthogonalised with respect to the main EVs. Session data were aggregated per participant using a second level fixed effects model. These 19 second level models were used to provide the input data for ROI analyses. Third level modelling was used to aggregate the data across participants in a 2x2 repeated measures ANOVA with Belief-Valence (B+/B-) and Desire-Valence (D+/D-) as within subjects factors. The final whole brain result was based on a mixed effects (ME) analysis with cluster based thresholding at $Z > 2.3$, $p_{corr} < 0.05$.

ToM Localizer Experiment Analysis

The localizer task was modelled as per Saxe and Kanwisher (2003). Statistical analysis was conducted using a GLM. Two EVs which reflected the two conditions, FB and FP, were convolved with a gamma-derived HRF. Second and third level modelling was used to aggregate the data across sessions and participants for the contrast of interest FB > FP. For examination of activation between the two experimental paradigms, post-stats processing of the group result was conducted as per the parameters used for the main belief-desire reasoning task (ME analysis, $Z > 2.3$, $p_{corr} < 0.05$).

Overlap Analysis

Using the whole brain data, any overlap between activations from the localizer task and the belief-desire task were identified using FSL's command line tools (fslmaths). A logical AND function was applied to the thresholded data ($Z > 2.3$, $p_{corr} < 0.05$) for the factors of Belief-Valence and Desire-Valence and the localizer FB > FP contrast.

ROI Analysis

ROI masks were created using the MarsBaR region of interest toolbox (version 0.42 marsbar.sourceforge.net) for SPM8 (www.fil.ion.ucl.ac.uk/spm). Masks comprised a sphere with a 5-mm radius centred on the single subject peak voxel within TPJ for the FB > FP localizer contrast. ROI analyses were carried out on each participant's aggregated sessional data for the 4 EVs modelled in the main belief-desire experiment. The mean percentage signal change (PSC) for each condition of interest within each ROI was extracted using FSL Featquery (www.fmrib.ox.ac.uk/fsl/feat5/featquery.html).

RESULTS

Belief-Desire Reasoning Task Behavioural Results

All reaction times (RTs) were recorded from the onset of the response probe. Any incorrect responses or data points that were 2 standard deviations outside of the participant's condition mean were removed for RT analysis. A 2x2 repeated measures ANOVA was conducted on the remaining data, with Belief-Valence (B+/B-) and Desire-Valence (D+/D-) as within subjects factors. This revealed significant main effects of Belief-Valence, where B- > B+ ($F(1,18) = 46.94$, $p < 0.001$, $\eta^2 = 0.72$) and Desire-Valence, where D- > D+ ($F(1,18) = 25.21$, $p < 0.001$, $\eta^2 = 0.58$) but no interaction ($F(1,18) = 0.21$, $p = 0.66$, $\eta^2 = 0.01$). Fig. 3A summarises the mean RT for correct responses given across the four conditions.

The participant's error rate was analysed in a further 2x2 repeated measures ANOVA. This also indicated significant main effects of Belief-Valence where B- > B+ ($F(1,18) = 22.55$, $p < 0.001$, $\eta^2 = 0.56$) and Desire-Valence where D- > D+ ($F(1,18) = 5.86$, $p = 0.03$, $\eta^2 = 0.25$), but no interaction between the two ($F(1,18) = 0.63$, $p = 0.44$, $\eta^2 = 0.03$). Fig. 3B illustrates the mean proportion of incorrect responses.

(A) **(B)**

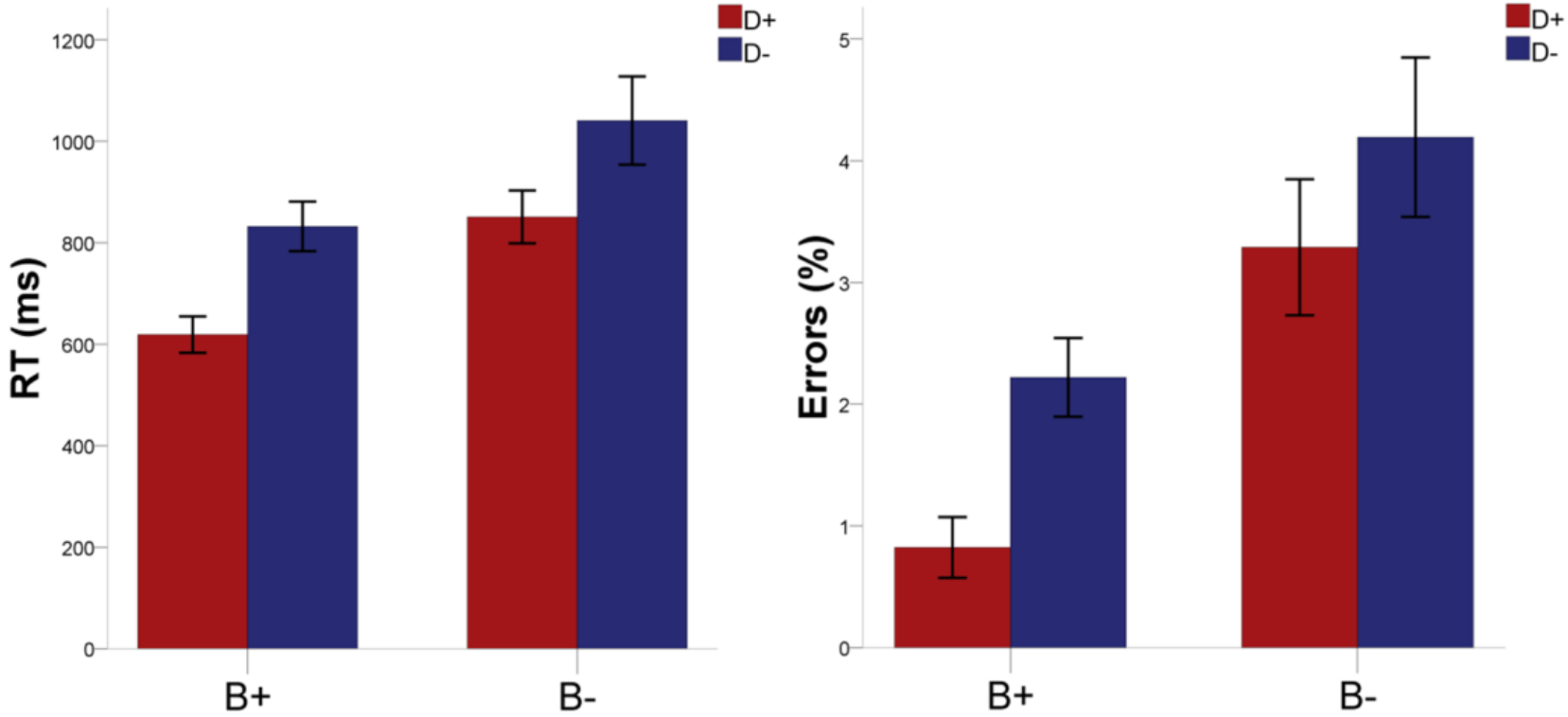

Fig. 3. Results: Belief-Desire Reasoning Behavioural Data
Error bars reflect +/-1 SE of the mean. (Panel A) Group mean reaction time per condition for correct responses (ms). (Panel B) Percentage of errors made per condition.

Whole Brain Analysis

Belief-Desire Reasoning Experiment

A 2x2 repeated measures ANOVA of the belief-desire reasoning task identified main effects of Belief-Valence (B+/B-) and Desire-Valence (D+/D-) but no interaction between the two factors. Manipulation of Belief-Valence recruited bilateral TPJ, superior parietal and occipital cortices, as well as frontal areas including the ACC (BA 32), bilateral dorsolateral prefrontal cortex (dlPFC) (BA 9, 46) and vlPFC including bilateral orbital frontal cortex, IFG and FO (BA 44, 45, 47) (Table 1; red shading Fig. 4). Varying Desire-Valence also elicited activation in bilateral TPJ, superior parietal and occipital cortices, and medial frontal regions including the ACC. However, in contrast to the factor of Belief-Valence, frontal activation was largely left lateralised, spanning both dlPFC and superior regions of vlPFC. Modulation of right frontal areas was limited to dlPFC (Table 2; green shading Fig. 4). Thus, whilst the valence of belief and desire both modulated activation in ACC, only belief was shown to influence the most inferior parts of vlPFC.

Table 1 Cluster Peaks for the Belief-Desire Reasoning Task: Factor of Belief-Valence

Hemisphere and region	Brodmann areas	Cluster size (voxels)	Peak MNI coordinates			Z-values
			x	y	z	
L inferior frontal gyrus, L middle frontal gyrus, L frontal operculum, L frontal orbital cortex	6, 8, 9, 38, 44, 45, 46, 47, 48	3134	-50	20	24	4.97
L temporoparietal junction, L supramarginal gyrus, L lateral occipital cortex	22, 39, 40	2859	-54	-52	26	4.92
R orbital frontal cortex, R frontal operculum, R inferior frontal gyrus, R middle frontal gyrus	9, 38, 44, 45, 46, 47, 48	1414	34	24	-6	4.85
R temporoparietal junction, R lateral occipital cortex, R middle temporal gyrus	22, 39, 40	1411	52	-54	24	4.78
L/R superior frontal gyrus, L/R paracingulate gyrus, L/R anterior cingulate cortex	8, 9, 24, 32	2069	0	28	46	4.37
R cerebellum crus I	-	561	18	-70	-34	3.78

Note. Clusters reflect results of 2-way repeated measures ANOVA for the factor of belief-valence (B+/B-). Table shows neural regions which are modulated by varying truth-status (true/false), $p_{corr} < 0.05$

Table 2 Cluster Peaks for the Belief-Desire Reasoning Task: Factor of Desire-Valence

Hemisphere and region	Brodmann areas	Cluster size (voxels)	Peak MNI coordinates			Z-values
			x	y	z	
L middle frontal gyrus, L inferior frontal gyrus	6, 9, 44, 45, 48	2339	-46	12	36	5.37
L/R precuneus	7	736	2	-66	42	4.63
L angular gyrus, L temporoparietal junction, L lateral occipital cortex	21, 39, 40	2141	-40	-56	52	4.46
L/R superior frontal gyrus, L/R paracingulate gyrus, L/R anterior cingulate cortex	8, 9, 24, 32	767	2	18	54	4.33
R angular gyrus, R temporoparietal junction, R supramarginal gyrus	7, 22, 40, 41, 48	1067	36	-48	42	3.80
R inferior frontal gyrus, R middle frontal gyrus	9, 44, 45, 48	467	44	28	20	3.69

Note. Clusters reflect results of 2-way repeated measures ANOVA for the factor of desire-valence (D+/D-). Table shows neural regions which are modulated by varying desire-status (approach/avoid), $p_{corr} < 0.05$

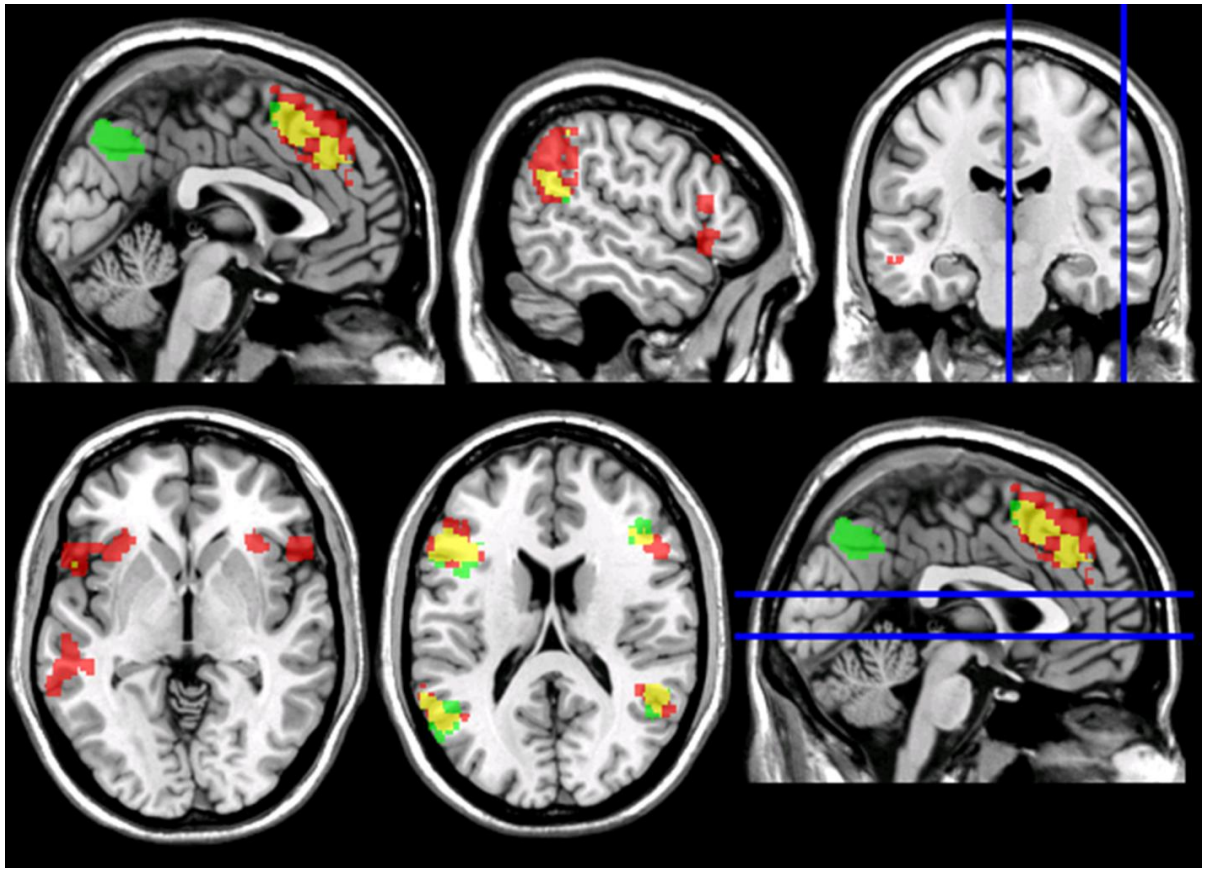

Fig. 4. Results: Belief-Desire Reasoning Whole Brain Analysis
Results from 2x2 repeated measures ANOVA whole brain analysis of the belief-desire reasoning task, with Belief-Valence (B+/B-) and Desire-Valence (D+/D-) as within-subjects factors. Selected slices highlight modulation in ToM and executive control regions for the factors of Belief-Valence (red) and Desire-Valence (green). Yellow areas indicate regions recruited by both factors (B/D). The group data are overlaid on the MNI brain template, showing significantly activated voxels where $Z > 2.3$, $p_{corr} < 0.05$. Slices from top left to bottom right, x = -1, 54; z = -2, 18 respectively. Images reflect Z-corrected F-stat images and are displayed in neurological convention, where left is represented on the left side of the image.

Theory of Mind Localizer Experiment

A mixed effects analyses of the whole brain localizer data identified neural regions that were more responsive to mental than physical representation (FB > FP, $p_{corr} < 0.05$). These results were consistent with previous ToM studies, showing that the FB > FP contrast recruits core regions of the ToM network, including bilateral TPJ and mPFC (Table 3; green shading Fig. 5A).

Overlap Analysis Results

Inspection of the activation maps from the group data suggested considerable overlap between neural regions recruited by the localizer task and the belief-desire reasoning task, as shown in Fig. 5A. An overlap analysis identified that only bilateral TPJ was required for all three variations of mentalizing (Fig. 5B).

Table 3 Cluster Peaks for the ToM Localizer Task, showing Activation where False Belief > False Photograph

Hemisphere and region	Brodmann areas	Cluster size (voxels)	Peak MNI coordinates			Z-values
			x	y	z	
R temporoparietal junction, R lateral occipital cortex, R middle temporal gyrus	21, 22, 39, 40, 42	4157	60	-58	18	4.76
L/R precuneus	7	2866	2	-58	36	5.56
L/R frontal pole, L/R medial prefrontal cortex, L/R superior frontal gyrus	8, 9, 10, 11	3511	-4	66	-12	4.65
L middle temporal gyrus	20, 21	1441	-58	-8	-20	4.72
L lateral occipital cortex, L temporoparietal junction, L angular gyrus	7, 19, 21, 39	1263	-42	-70	-38	4.08
L cerebellum crus II	-	957	-30	-80	-40	3.84
L cerebellum IX	-	382	-4	-56	-46	3.92

Note. Clusters reflect results from t-test of FB>FP. Table shows neural regions which are more responsive to false-belief than false-photo stimuli, $p_{corr} < 0.05$

(A) **(B)**

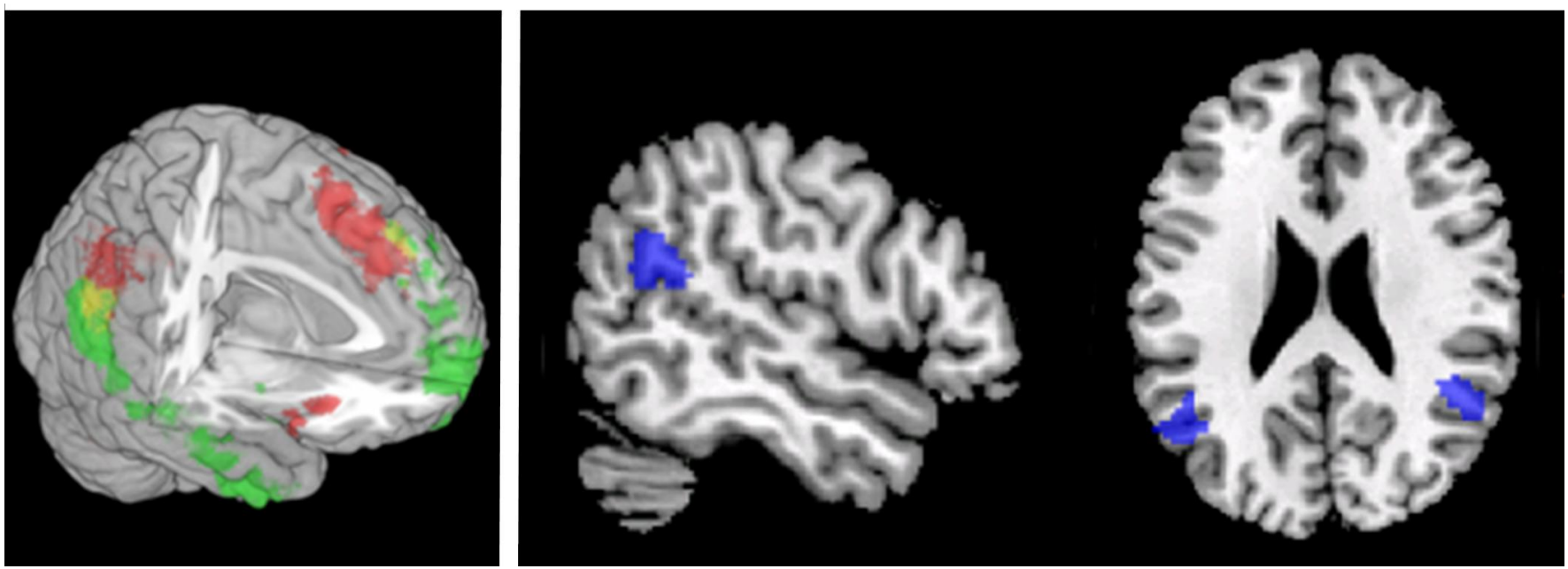

Fig. 5. Results: Belief-Desire Reasoning & Localizer Task Overlap Analysis
(Panel A) Activation map for the contrast FB > FP (green) shown with the cluster maps from the belief-desire reasoning task, where the factors of Belief-Valence and Desire-Valence are represented by a single colour (red). Yellow areas indicate regions recruited for both the localizer and the belief-desire reasoning task. Each map is overlaid onto the MNI brain template and shows significantly activated voxels where Z > 2.3, $p_{corr} < 0.05$. **(Panel B)** Blue clusters reflect conjunction between localizer contrast FB > FP, and Belief-Valence Desire-Valence factors B+/B- & D+/D-, $p_{corr} < 0.05$. Slices x = 52, z = 24. Images reflect Z-corrected t-stat images and are displayed in neurological convention, where left is represented on the left side of the image.

ROI Results

As bilateral TPJ were the only regions identified for both mental representation (localizer) and variation in mental state valence (belief-desire task), we focused ROI analyses on these areas. ROIs were identified using the localizer task in 18 of 19 individual participants in the rTPJ and 18/19 in lTPJ. ROI analysis was conducted on data from the belief-desire reasoning task and a 2x2 repeated measures ANOVA conducted on the mean PSC data for each ROI (Fig. 6). The right TPJ's response was higher when reasoning about a false than a true belief ($F(1,17) = 20.43$, $p < 0.01$, $\eta^2 = 0.55$) and higher for avoidance versus approach desire ($F(1,17) = 9.47$, $p < 0.01$, $\eta^2 = 0.36$). No interaction existed ($F(1,17) = 2.87$, $p = 0.11$, $\eta^2 = 0.14$). Similar effects were detected in lTPJ where it's response was higher when reasoning about a false- than a true-belief ($F(1,17) = 12.73$, $p < 0.01$, $\eta 2 = 0.43$) and higher for avoidance- versus approach-desire ($F(1,17) = 28.64$, $p < 0.001$, $\eta 2 = 0.63$), but no interaction existed between Belief- and Desire-Valence ($F(1,17) = 1.97$, $p = 0.18$, $\eta 2 = 0.10$).

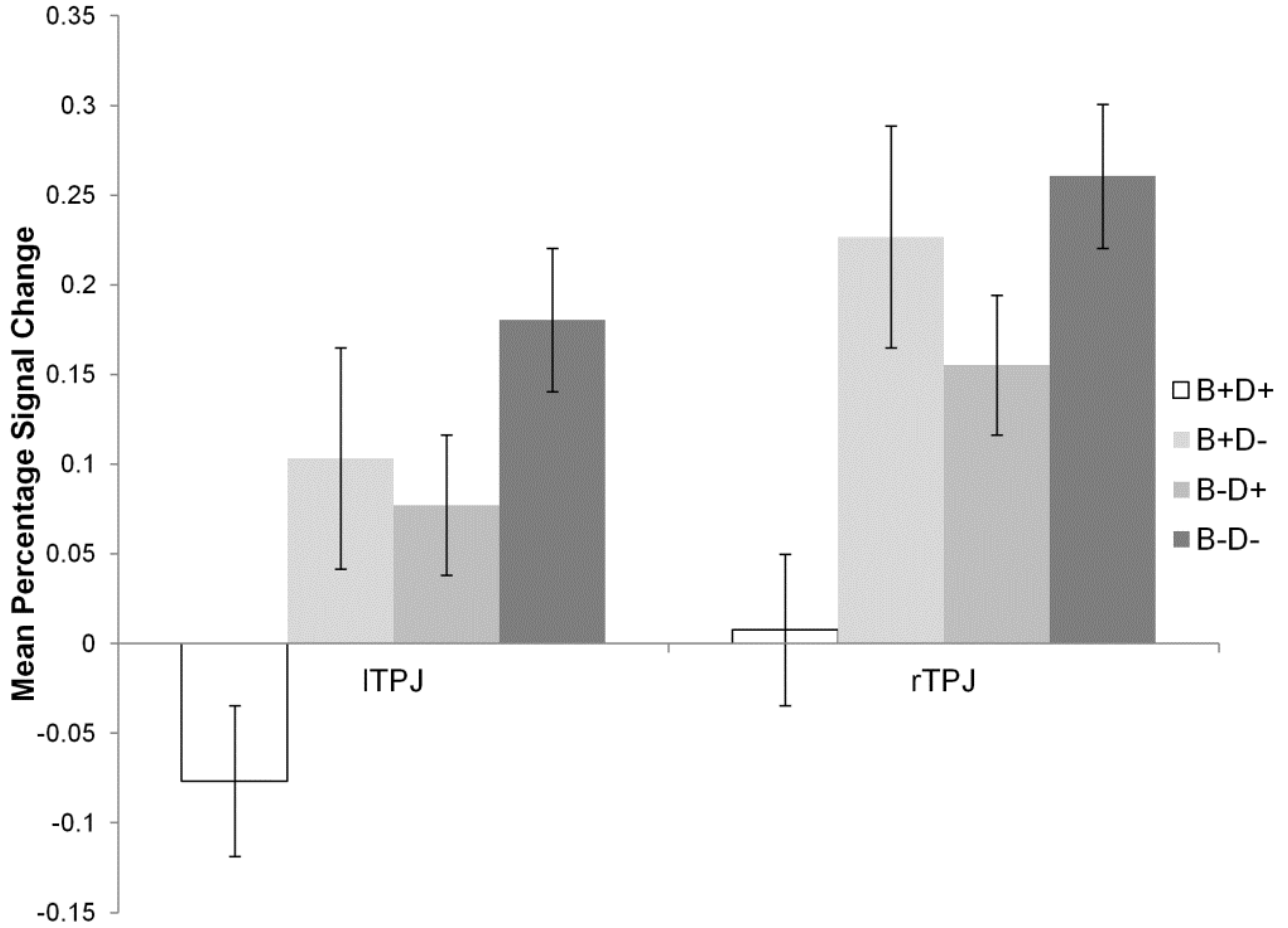

Fig. 6. Results: Belief-Desire Reasoning Region of Interest Analysis
Error bars reflect +/-1 SE of the mean. Results from the ROI analysis, where ROI masks generated using the localizer task were applied to the belief-desire task. Group mean percentage signal change (PSC) per condition

DISCUSSION

Behavioural evidence suggests that negatively valenced mental states – false beliefs and avoidance desires – are more difficult to process than their positively valenced counterparts. On developmentally sensitive tasks, young children pass false belief and avoidance desire tasks at a later age than true belief and approach desire tasks (Cassidy, 1998). Suitably adapted tasks demonstrate that adult participants, too, show a similar pattern of relative difficulty, reflected in response times and residual error rates (Apperly et al., 2011; German & Hehman, 2006). Moreover, in both children and adults, performance on such tasks is associated with independent tests of EC (e.g., Carlson & Moses, 2001; German & Hehman, 2006). The neuroimaging literature consistently identifies TPJ and mPFC as core ToM regions, but less is known about how activity in these regions is modulated by psychologically relevant differences between positive and negative valence. Likewise, little is known about how and when neurocognitive systems for EC are recruited in the service of different aspects of ToM. We addressed these issues in the current study by manipulating the valence of belief and desire states and by examining neural activity during the response phase of each trial, during which the behavioural costs of belief-desire reasoning have been observed on this task.

Do our factors of Belief-Valence and Desire-Valence recruit any regions of the ToM network?

We set out to investigate how variation in the valence of belief and desire states affects recruitment of the ToM network. A whole brain analysis demonstrated that variation in mental state valence modulates activity in neural regions regularly implicated in general ToM tasks including temporoparietal, medial parietal and some prefrontal regions. This finding

converges with evidence from a small number of studies that suggest that these regions not only respond to ToM tasks in contrast to non-ToM baseline tasks, but also that their activity varies according to the valence of belief and/or desire (Abraham et al., 2010; Sommer et al., 2007; van der Meer et al., 2011). Importantly, we find these effects during a canonical ToM task that requires participants to predict the action of an agent on the basis of belief and desire.

Alongside TPJ, anterior rostral areas of the mPFC are also commonly implicated in studies of ToM (Amodio & Frith, 2006; Carrington & Bailey, 2009; Lieberman, 2007; Mar, 2011; Van Overwalle, 2009). Whilst belief and desire reasoning modulated dorsal areas of the medial frontal cortex – particularly dorsal ACC – our novel paradigm showed no activation in anterior rostral mPFC. This finding contrasts with the ToM localizer task, which did show activity in anterior rostral mPFC. We believe that this pattern may be understood on the hypothesis that rostral mPFC is recruited for ToM to the degree that participants must go beyond the information immediately available to them, making social inferences about traits and norms, engaging in self-reflection or episodic thinking about the past or future (Gilbert, Spengler, et al., 2006). Such requirements are common in laboratory tasks and in everyday ToM, but are not a necessary feature of ToM cognition. In our belief-desire task participants were directly informed of the character's mental states, and the correct prediction of his or her action was wholly determined by deductive reasoning from this information. Thus, although participants needed to represent and reason about mental states, there was simply no need for inferences about traits, self reflection or episodic thinking. In contrast, the localizer task involved vignettes that, though short, did require participants to construct a situational context in which the character's mental states might be inferred. We suggest that it may be this need

for elaborative processing that results in the recruitment of rostral mPFC in the service of ToM inferences.

If our factors of Belief-Valence and Desire-Valence recruit regions of the ToM network, is this just because those regions are involved in attention/executive control, not because they are involved in ToM per se?

There are three main alternate explanations as to the role of TPJ in ToM. One is that this region responds specifically to transient mental states, regardless of their content or status; thus, TPJ may be specialised towards ToM (e.g., Saxe & Kanwisher, 2003; Van Overwalle, 2009). Support for this theory is found in data which pinpoint TPJ for a variety of ToM, but not control, tasks. This includes the attribution of beliefs (e.g., Aichhorn et al., 2009; Saxe & Kanwisher, 2003; Saxe & Wexler, 2005; Scholz et al., 2009) and, although little explored, desires (Saxe & Kanwisher, 2003). The second possibility is that TPJ may regulate the distinction between self and other (e.g., Brass, Ruby, & Spengler, 2009; Decety & Lamm, 2007). Activation of TPJ is a consistent feature of both mentalizing and seemingly disparate tasks such as the inhibition of imitative behaviour. It has therefore been suggested that TPJ is recruited for situations which require a person to disengage self from other, so that an individual can appropriately assign behaviours or mental states as belonging to an external agent. Lastly, it has been suggested that TPJ activation is observed in ToM tasks because TPJ supports domain-general processes that are unintended confounds of ToM tasks, such as reorienting spatial attention away from miscued locations (e.g., Mitchell, 2007; Rothmayr et al., 2011). When applied, for example, to a false belief scenario, this process might reflect the need to redirect one's attention from location A ("true" location) to location B ("false" location). It is suggested that, as a result, ToM and exogenous attention tasks mutually

activate right TPJ, which indicates that there may be some shared attentional component between ToM and spatial reorienting (Mitchell, 2007; Rothmayr et al., 2011).

Our findings do not fit well with the last of these three possibilities. The localizer task subtracted activation observed during false belief trials (which involve reasoning about false beliefs and management of attention between "false" and "true" locations) from activation observed during false photograph trials (which involve reasoning about photographs that are outdated/false and management of attention between "false" and "true" locations). Since the need to manage attention between "false" and "true" locations is present in both the false belief and false photograph conditions, and indeed, appears present to a similar degree, little activation due to such attention management is likely to survive the subtraction between these conditions. Instead, the surviving activation is more likely to be due to a difference between reasoning about false beliefs compared with false photographs. It is noteworthy, then, that this surviving activation in bilateral TPJ overlaps substantially with regions modulated by our novel belief-desire task. We think it unlikely that the common activation across these comparisons is due to a confounding requirement to reorient attention that has nothing to do with ToM.

Our findings also pose a challenge for the claim that TPJ is specialised for ToM and responds specifically to such transient mental states, regardless of their content or status (e.g., Saxe & Kanwisher, 2003; Van Overwalle, 2009), because we found that activity in these regions was modulated by the valence of both beliefs and desires. However, our findings might be reconciled with this theoretical interpretation by supposing that TPJ is playing a similar functional role across these conditions, but its activity is up- or down-regulated by the relative difficulty of the different belief-desire conditions. The participants in the present study were slower to respond to both false belief and avoidance desire scenarios, and for this

reason alone, activity in TPJ may have been held high for longer, or held higher overall. A further possibility is that TPJ is playing distinct functional roles across our belief and desire conditions, due to differential demands of representing true versus false beliefs and approach versus avoidance desires, or of making action predictions on the basis of this information. One potential source of differential demands is the need to maintain a distinction between self and other (e.g., Brass et al., 2009; Decety & Lamm, 2007), though this need varies much more obviously between true and false beliefs than between positive and negative desires. What is potentially interesting in this general interpretation is that it offers a way of combining the insights of the other two: on the one hand TPJ recruitment during ToM tasks may not be due to confounding demands on attentional control in ToM tasks, on the other it may be that attentional control is intrinsic to ToM problems, not least in order to maintain and switch between representations of self and other.

Do we observe differential activation of EC regions due to the Belief-Valence factor compared with the Desire-Valence factor?

Leslie and colleagues find that false belief and avoidance desire will attract greater processing costs than true belief and avoidance desire (Friedman & Leslie, 2004; Friedman & Leslie, 2005; Leslie et al., 2005; Leslie & Polizzi, 1998). Our data converge with these findings and the wider literature on behavioural performance in adult belief-desire reasoning (Apperly et al., 2011; German & Hehman, 2006). Leslie and colleagues additionally specify that belief and desire reasoning is supported by a common process (termed a 'selection processor' in their account) which directs executive selection resources in attentionally demanding situations, for example, when attributing negatively valenced mental states. Our data are consistent with this idea and identify ACC as a possible candidate for EC processes

associated with such variation in task difficulty. Whilst most extensively examined in the cognitive literature (e.g., Botvinick, Cohen, & Carter, 2004; Botvinick, Nystrom, Fissell, Carter, & Cohen, 1999; Carter et al., 1998), ACC is increasingly acknowledged to play an important role in supporting social cognition (Amodio & Frith, 2006; Lieberman, 2007). Converging electrophysiological and neuroimaging data suggest a functional division within ACC, where dorsal areas subserve conflict monitoring and error detection, and rostral-ventral areas are primarily involved in the assessment of motivational or emotional information (Amodio & Frith, 2006; Bush, Luu, & Posner, 2000; Devinsky, Morrell, & Vogt, 1995). For the present experiment, both the valence of both belief and desire states was shown to modulate activation in dorsal ACC, suggesting that reasoning about very basic belief and desire states draws on a common cognitive process. As seen in our behavioural data, manipulation of mental state valence yielded processing costs in terms of error rates and response latencies. On this basis we propose that dorsal ACC indexes conflict (between self and other perspectives, and between the agent's belief about the object and his desire to avoid it) in order that further executive processes, such as inhibition and selection, may be initiated.

We have suggested that increased attentional demands may help explain behavioural difficulty with negatively valenced mental states, but it may be that this does not exhaust the role of EC in ToM. As described in the introduction, a growing body of research suggests that participants will be slower and more error prone when holding in mind mental states which are incongruent with their own self perspective, such as a false belief or conflicting (not merely avoidance) desire state (Ruby & Decety, 2003; Samson et al., 2005; van der Meer et al., 2011). In the present study, we manipulated congruence with self other perspectives by asking participants to make predictions about a protagonist's behaviour in true and false belief scenarios. In contrast, our manipulation of approach versus avoidance desire did not result in

differences in congruence of self and other perspectives, and so did not vary the need for self-perspective inhibition.

Belief-Valence, but not Desire-Valence, was seen to recruit the most inferior parts of bilateral vlPFC. Variation in the conflict between the perspectives of the participant and of the agent was manipulated in the Belief-Valence, but not Desire-Valence, condition. Thus, our data are consistent with the view that activation in vlPFC is modulated by variation in the need for self perspective inhibition, and show that this is a critical difference between true and false belief trials, as well as between false belief trials between which the salience of self-perspective is experimentally varied (Samson et al., 2005; van der Meer et al., 2011). The present dataset therefore provides strong evidence for a distinct role for EC beyond the generic control of attention during ToM tasks. In addition, EC is necessary when a perspective difference between self and other exists, as is the case for false belief. This converges with behavioural data from the current study and others indicating that knowledge of the true state of affairs interferes with the ability to select the believed (i.e. false) location, when the real and believed location are incongruent, giving rise to the well-known phenomenon of egocentric biases and errors (Bernstein et al., 2004; Birch & Bloom, 2004; Birch & Bloom, 2007; Mitchell & Lacohee, 1991). The process of inhibiting this self perspective, we suggest, specifically recruits vlPFC. Importantly, such activity would necessarily be missed in studies using the best-controlled comparisons between ToM and non-ToM tasks. For example, it would not be observed in the Saxe and Kanwisher (2003) ToM localizer because both the false belief and the false photograph conditions require inhibition of self perspective, and so any associated activation would be lost in the subtraction of one condition from another.

CONCLUSION

The present study provides evidence that converges with and extends a number of findings concerning the functional and neural basis of ToM. We find evidence that activation in TPJ is modulated by the valence of mental states, suggesting that this region is not responsive to transient mental states per se (e.g., Saxe & Kanwisher, 2003; Van Overwalle, 2009), but rather the content of such mental states. We find evidence that the mere requirement to represent a mental state may be insufficient to recruit rostral mPFC, but that this region is recruited when mental states need to be inferred on the basis of contextual information, consistent with Amodio and Frith (2006); Van Overwalle (2009).We also find evidence of the recruitment of neural regions associated with EC, which converges with behavioural evidence that ToM problems often require domain-general EC processes, as well as processes that might be more specific to ToM (Apperly et al., 2008; Apperly et al., 2011; Carlson & Moses, 2001; Carlson et al., 2002; Carlson et al., 1998; Cassidy, 1998; Friedman & Leslie, 2004; German & Hehman, 2006; Leslie et al., 2005; Leslie & Polizzi, 1998; Perner & Lang, 1999).

The present study significantly extends understanding of the relationship between ToM and EC, and the neural systems that support these abilities. ToM problems that participants find more difficult to solve – such as those involving false belief and avoidance desire – result in greater activity in neural systems involved in attentional control, such as ACC, and also in parts of the "ToM network", such as TPJ. Importantly, this effect of general difficulty can be distinguished from a more specific effect due to the need to resist interference from self perspective. This need only arises when there is a perspective difference between self and other – as in the false belief condition of the current study – and appears to recruit vlPFC in a distinctive manner. Nonetheless, additional work is required to

further examine the role of EC in ToM and, in particular, the involvement of vlPFC in inhibition of self-perspective. The use of an established EC paradigm in parallel with a tightly controlled ToM task, such as was presented here, would advance our understanding of the neural basis of those domain-general processes that support ToM. Moreover, specific manipulations in terms of desire reasoning, where an agent's desire state is made systematically congruent or incongruent with self, would serve to further delineate the role of vlPFC in inhibition of self-perspective.

In sum, we demonstrate how the virtues of subtractive, "localizer" methods and methods that allow psychologically relevant parameters to be varied orthogonally may be combined to give a deeper understanding of the cognitive and neural basis of ToM than would be possible with either method alone.

CHAPTER 3:

REPRESENTATION, CONTROL OR REASONING? DISTINCT FUNCTIONS FOR THEORY OF MIND WITHIN THE MEDIAL PREFRONTAL CORTEX[5]

[5] This chapter is published: Hartwright, C. E., Apperly, I. A., & Hansen, P. C. (2013). Representation, Control, or Reasoning? Distinct Functions for Theory of Mind within the Medial Prefrontal Cortex. *Journal of Cognitive Neuroscience*. doi: 10.1162/jocn_a_00520

ABSTRACT

The medial prefrontal cortex (mPFC) is frequently reported to play a central role in Theory of Mind (ToM). However, the contribution of this large cortical region in ToM is not well-understood. Combining a novel behavioural task with fMRI, we sought to demonstrate functional divisions between dorsal and rostral mPFC. All conditions of the task required the representation of mental states (beliefs and desires). The level of demands on cognitive control (high versus low) and the nature of the demands on reasoning (deductive versus abductive) were varied orthogonally between conditions. Activation in dorsal mPFC was modulated by the need for control, whereas rostral mPFC was modulated by reasoning demands. These findings fit with previously suggested domain-general functions for different parts of mPFC, and suggest that these functions are recruited selectively in the service of ToM.

INTRODUCTION

Theory of Mind (ToM) is a term used to describe the ability to attribute mental states such as beliefs, desires and intentions to other individuals. By applying a ToM, social agents are better able to predict the behaviour of those around them, and may additionally direct our own behaviour in terms of whether we choose to deceive, cooperate, or empathise with others (Gallagher & Frith, 2003). This ability to 'mentalize' has received much attention from the neuroimaging community over the last decade, and has identified a set of brain regions that are consistently responsive when thinking about the contents of other people's minds: the left and right temporoparietal junction (TPJ), medial parietal cortices including the precuneus and posterior cingulate, the temporal poles and the medial prefrontal cortex (mPFC) (for reviews

see Carrington & Bailey, 2009; Lieberman, 2007; Mar, 2011; Van Overwalle, 2009). The most prominent debate within the literature, however, surrounds how medial prefrontal and temporoparietal regions support a functioning ToM.

One challenge for social neuroscientists is to localise ToM processes more precisely by identifying functional subdivisions within the anatomical regions associated with ToM. mPFC, in particular, comprises a large area of the cortex and is involved in many aspects of social cognition (Amodio & Frith, 2006), together with an array of executive processes such as reallocation of attention, action monitoring and control (Lieberman, 2007; Ramnani & Owen, 2004; Rushworth, Buckley, Behrens, Walton, & Bannerman, 2007), relational integration and multitasking (Gilbert, Spengler, et al., 2006; Ramnani & Owen, 2004), outcome monitoring (Gilbert, Spengler, et al., 2006), working- and episodic-memory (Gilbert, Spengler, et al., 2006; Lieberman, 2007; Ramnani & Owen, 2004; Spreng et al., 2009) and default mode or spontaneous 'at rest' cognition (Amodio & Frith, 2006; Ramnani & Owen, 2004; Spreng et al., 2009). Since ToM is a social process but undoubtedly also entails executive processing, attention and reasoning (Apperly, 2011), it is perhaps unsurprising that the role of mPFC in ToM remains unclear (Rothmayr et al., 2011).

ToM: A Task Analysis

It is likely that activation of mPFC in some ToM tasks reflects executive processes that are an incidental feature of the task used to present the ToM problem, and thus do not constitute core processes which underlie ToM. Nonetheless, there are also good reasons for believing that specific sub regions of mPFC are more centrally involved in ToM. A task analysis of ToM suggests three processes which may explain how specific regions of mPFC are involved in mentalizing. First, a common theme across all forms of ToM reasoning is the

requirement for *representation* of a mental state. Thus, regardless of whether an individual is asked to reason about an agent's belief, desire, intention, or the like, it is necessary to represent a mental state of some kind. The frequency with which more rostral areas of mPFC are recruited for ToM and other social cognitive functions has lead researchers to tentatively suggest that mPFC might subserve such a process (Amodio & Frith, 2006; Frith & Frith, 2006; Frith & Frith, 2003). Other data suggest that TPJ may be even more selectively responsive than mPFC to representation of mental states (Aichhorn et al., 2009; Saxe & Kanwisher, 2003; Saxe & Wexler, 2005; Scholz et al., 2009). Resolution of this debate is unnecessary for our current purposes. What matters for now is that there are grounds to suppose that mPFC may be involved in representation of mental states, and that it is possible to distinguish this representational requirement, which attends all ToM tasks, from other important requirements for cognitive control and reasoning, which vary across ToM tasks or experimental conditions.

Second, a large body of behavioural and neural evidence indicates that ToM is associated with processes for cognitive *control* (e.g., Apperly, 2011; Lieberman, 2007). Control processes for inhibition, conflict monitoring and working memory are not only necessary for meeting the demands of the relatively complex stories or cartoons frequently used to study ToM, but also seem to be essential for ToM *per se*. For example, in the classic false belief paradigm (see Wimmer & Perner, 1983), an agent holds an outdated, or 'false', belief about reality. Predicting the agent's action requires participants first to infer that the agent's belief is different from their own, second to hold this false belief in mind and not confuse it with their own knowledge, and third to predict the agent's action selectively on the basis of the agent's belief, rather than according to the participant's own knowledge of the right course of action. Behavioural data from both children and adults suggest that the effort

required for ToM reasoning (as indexed by response times and error rates) depends upon whether an agent's belief is true or false, and whether their desire is to approach or avoid a target object (Apperly et al., 2011; German & Hehman, 2006; Hartwright et al., 2012). Attempts to understand the neural basis of such effort consistently identify more dorsal regions of mPFC (dmPFC) approximating Brodmann Areas (BA) 8, 9 and 32. For example, dmPFC is modulated by contrasting ToM concepts where maximal conflict exists, as is the case in false belief reasoning versus reasoning about an agent whose belief is a 'true' representation of reality (Döhnel et al., 2012; Hartwright et al., 2012; Sommer et al., 2007).

Importantly, recent evidence suggests that the contribution of frontal regions does not just vary according to overall task difficulty, but according to the source of that difficulty in the ToM task. In a recent study on which the present paradigm is based, participants predicted the action of an agent whose belief was either true or false, and whose desire was either to approach or avoid an object (Hartwright et al., 2012). Both factors have the potential to vary cognitive conflict, because both false belief and avoidance desire lead the agent into counter-intuitive actions away from a salient target object. However, only the belief factor (true versus false) leads to systematic variation in perspective between the character and the participant. In this study dmPFC was modulated equally by the belief and desire factors, suggesting that it was performing a general role in resolving cognitive conflict. This contrasted with more lateral prefrontal regions, such as bilateral inferior frontal gyrus (IFG), which responded differentially to true- versus false-belief, but not to approach- versus avoidance-desire. These findings suggest that dmPFC underlies frontline control processes which monitor conflict during ToM reasoning, whereas other regions, such as IFG, are recruited for more specific processes such as inhibition of self-perspective. This theory about the contribution of dmPFC converges with neuroimaging research outside of the social domain which identifies mPFC,

particularly more dorsal regions including the dorsal anterior cingulate cortex (dACC), in conflict monitoring and error detection. It has been shown that activation in dmPFC is modulated by task difficulty, where those tasks that attract increased error-rates and response latencies make the most demands on this region (Botvinick et al., 2004; Botvinick et al., 1999; Bush et al., 2000).

The final process we propose within our task analysis of ToM is the focus of the current study, and refers to the different roles of *reasoning*. Philosophers, Logicians and Computational Scientists have long debated the formulation of reasoning. These debates are beyond the scope of the present paper; however, we borrow two theoretical concepts in order to illustrate how different approaches to ToM can activate alternative modes of inference and their neural correlates. In the belief-desire task used by Hartwright et al. (2012), participants were told three facts for each trial: the agent's belief about the location of an object, the agent's desire to seek out or avoid the object and the true location of the object. Given this information, participants had to identify which location the agent would choose on the basis of his belief and desire state. Thus, participants had to reason "deductively", so no reasoning beyond the facts explicitly presented was required (Morris, 1992; Pagnucco, 1996). Unlike the vast majority of neuroimaging studies of ToM, activation within rostral mPFC (rmPFC), approximating BA10, was noticeably absent from this deductive ToM paradigm.

However, many ToM studies, and certainly a good deal of ToM outside of the laboratory, do not provide explicit access to all of the facts necessary to solve the task (Jenkins & Mitchell, 2009). Consequently, the individual is required to engage in open-ended "abductive" reasoning about an agent's behaviour in order to use their ToM effectively. Consider a typical ToM vignette taken from Saxe and Andrews-Hanna (n.d.),

The morning of the high school disco Sarah placed her high heel shoes under her dress and then went shopping. That afternoon, her sister borrowed the shoes and later put them under Sarah's bed.

-

Sarah gets ready assuming her shoes are under her dress.

True *False*

Unlike the deductive approach, reasoning here is used to explain an observation on the basis of a hypothesis, which may or may not turn out to be correct. Here, participants are required to reason abductively – that is to infer the most likely cause (Sarah's belief that her shoes are under her dress), on the basis of the given effect (that Sarah gets ready unaware that her shoes might not be where she expects to find them) and a ToM principle (that Sarah will look for the shoes on the basis of her belief state). Here, then, reasoning is an inference to the most appropriate explanation (Menzies, 1996). Reasoning deductively, where one uses a set of rules and preconditions to generate a conclusion (Menzies, 1996), is likely to involve cognitive process that differ from an abductive approach involving reasoning to explain an observation (Morris, 1992). Whilst the neural basis of deductive reasoning has been studied extensively (see Prado, Chadha, & Booth, 2011 for a recent review), little work has been done for abductive inference. Nonetheless, when considered in terms of the underlying process of thinking beyond the given information, studies indicate that rmPFC is recruited when participants are required to reason beyond the constraints of the information immediately available to them (Gilbert et al., 2007; Hartwright et al., 2012; Jenkins & Mitchell, 2009), whether the context is social or non-social. This leads to the hypothesis that, rather than being involved in *representing* mental states, rmPFC is recruited whenever ToM tasks require

abductive reasoning. This would account for the frequent observation of rmPFC activation because abductive reasoning is very common in ToM tasks.

However, there is an alternative explanation for the lack of variable mPFC activation in Hartwright et al. (2012), that remains consistent with the hypothesis that rmPFC supports the representational demands of ToM, as touched upon earlier in our task analysis. The need to represent mental states was present across all conditions in Hartwright et al.'s deductive task; consequently, rmPFC might not be identified by orthogonal comparisons across conditions if this region generally services the process of representation. Therefore, the present study manipulates the need for abductive reasoning within-task in order to disambiguate these two possibilities.

In sum, there are multiple theoretical reasons for thinking that mPFC might be involved in ToM, and several competing hypotheses designed to account for this. Furthermore, there are grounds for thinking that there might actually be functional differentiation within mPFC, which a number of researchers have suggested would be best identified using a single, within-experiment, within-subjects design (Abu-Akel & Shamay-Tsoory, 2011; Carrington & Bailey, 2009). The task analysis presented here proposes three separate processes for ToM: representation, control and reasoning. *Representation*, we argue, is a ubiquitous feature of mentalizing. *Control* and *reasoning* processes, conversely, vary across different ToM tasks. The latter two ToM processes lend themselves well to manipulation within a single, repeated-measures paradigm. Consequently, the present study served two purposes. First, to replicate Hartwright et al.'s earlier finding that dmPFC is modulated as a function of control. Second, by making a minimal change to our previous paradigm, we aimed to demonstrate that we could recruit the previously absent rmPFC by including a condition which required abductive reasoning. In order to achieve this, we present

a 2x3 repeated measures orthogonal design. The valence of an agent's belief was either true or false; the valence of desire was either approach- or avoidance- (as in Hartwright et al., 2012), or it was unspecified. The novel, unspecified, condition required participants to reason about whether they thought the agent would have an approach- or an avoidance-desire, on the basis of what sort of person they thought the agent was. We expected those mental states where conflict is inherent, but presented unambiguously in our paradigm (i.e. false belief, avoidance desire), to preferentially recruit dmPFC. Conversely, a mental state that required abductive reasoning (i.e. desire unspecified), was expected to preferentially activate rmPFC.

METHOD

Participants

Twenty right-handed adults participated in the fMRI experiments (12 female; mean age = 21 years). All were native-English speakers and were given a small honorarium for their participation. The study had research ethics approval from the University of Birmingham. All participants gave written consent to participate in the study.

Materials and Procedure

Pre-screen

A pre-screen to determine suitability to participate was conducted several days before collecting any neuroimaging data. This consisted of a handedness measure, using a modified form of the Annett Handedness Questionnaire (1970), and a reading scale – the Wide Range Achievement Test Third Edition (WRAT-3) – to ensure reading proficiency commensurate with the experimental tasks.

Participants were informed that the social judgements task required them to make predictions about how real individuals played a game in a previous experiment. They then completed a computer based interactive training session and two test blocks of the social judgements experiment. Those who performed above chance[6] on the test blocks were invited to participate in the fMRI experiment (see Appendix 1 for detailed participant screening information). Of twenty six prospective participants, 6 individuals (4 female; age range 18-24, $\overline{X}$ age = 21 years) were unable to perform the social judgements experiment to above chance at the pre-screen. These individuals were not invited to participate in the fMRI experiment.

Social Judgements Experiment

The social judgments task was based on a paradigm devised by Apperly et al. (2011) and Hartwright et al. (2012). The experiment comprised an orthogonal design where a protagonist's belief state (true (B+) or false (B-)) and desire state (approach (D+), avoid (D-) or unspecified (D±)) was systematically manipulated, resulting in six equally occurring conditions: B+D+, B+D-, B+D±, B-D+, B-D-, B-D±. Immediately prior to collecting any neuroimaging data, participants were again informed that the task was based on real game playing data from real individuals, and that the participant's job was to predict how these individuals played the game. All participants then revisited the interactive training program used in the pre-screen and completed a further practice block outside of the MRI scanner. Note that none of the practice trials were used in the fMRI experiment.

The fMRI experiment required the participants to watch and predict which one of two different coloured boxes a character, referred to as 'the contestant', would open in a virtual game show. A single round (i.e. trial) of the game show consisted of the contestant being told

[6] Chance at $p < 0.05$, based on a binomial distribution

what prize was on offer, followed by them guessing which one of the two boxes contained the prize, ending in them opening one box. The contestant would win whatever was in the box they opened; however, one box was always empty and the other always contained the prize. If the box contained a prize, they would win it. If it was empty, they would win nothing and play a new round.

Each contestant played multiple rounds. The prizes ranged in desirability and, as the contestant could only win a finite number of prizes, it was not always in their interest to play to win every prize. If the contestant liked the prize on offer, they would open the box where they guessed the prize was hidden, in the hope of winning that prize. If they did not like the prize, they would open the opposite box to where they guessed the prize was hidden (i.e. the empty box), in the hope of having another chance to win something more to their liking. Note, however, that the game show was designed such that the contestant would only take home a prize in half of the trials. Furthermore, in half of the 'winning' trials, the contestant would win a prize that they did not actually want to win.

Whilst the fMRI data were collected, participants watched a computer based mock-up of the contestants playing the afore-described game show. The participants' job was to predict which box the contestant opened on the basis of the contestant's belief and desire state, in terms of which of the two boxes the contestant believed contained the prize, and the contestant's desire to win or gamble and play on for a better prize. Participants were always told the contestant's belief about the location of the prize and the true location of the prize, but had to infer the contestant's desire to win the prize, based upon a colour photograph which depicted the contestant smiling (D+), frowning (D-) or with a neutral (D±) expression. The training sessions conducted prior to collecting any fMRI data taught the participants to treat a smiling face as signalling the contestant's pleasure and, therefore, their desire to open the box

that they thought contained the prize (approach desire), and a frowning face as signalling the contestant's displeasure and, therefore, their desire to avoid opening the box that they thought contained the prize (avoidance desire). Where the contestant was shown with a neutral expression, participants were asked to consider what sort of person they thought the contestant was, in terms of what their likes and dislikes might be, and to select which box they thought the contestant would open (unspecified desire). Just as with approach/avoid (D+/-) trials, in these unspecified desire (D±) trials, participants were told to select the box that the contestant believed contained the prize if they thought the contestant would have played to win the prize on offer, or to select the opposite box (i.e. what the contestant believed to be the empty box) if they thought the contestant would have wanted to avoid winning the prize. Note that in all cases, participants were told to make their responses on the basis of the contestant's belief state, which could be either true (B+) or false (B-), therefore requiring them to ignore their own knowledge of the true location of the prize.

Each block of trials opened with an instruction screen followed by an initial interstimulus interval (ISI) of 11600 ms. A single trial comprised three centre justified statements shown singularly for 1600 ms and separated by a 500 ms fixation period, followed by a picture response probe shown for 2500 ms, then a rest period. A variable ISI was used for rest (range = 9000-14000 ms, $\overline{X}$ = 11500 ms) during which a small fixation dot was displayed. Each trial lasted 8800 ms, excluding fixation. The experiment comprised 6 separate blocks, each of which contained 28 trials and took 9 m 36 s to complete.

Each trial opened with a prize statement (e.g., The prize on offer: designer shoes), followed by either a belief statement (e.g., The contestant thinks the prize is in the red box) or a reality statement (e.g., The prize is in the blue box), then the remaining belief or reality statement. The temporal order of belief and reality statements was randomised, but contained

an equal number of each ordering overall. The final statement was followed by a response probe then rest. Participants were able to respond from the onset of the response probe, using a two button box placed in their right hand. Participants responded by pressing the left button to indicate the left prize box and the right button to indicate the right prize box.

Two formats of response probe were used. The format indicated to the participant what type of response to give. If a full colour photograph of the contestant was shown, the participant had been trained to indicate which box they thought the contestant opened, based upon the contestant's belief-desire state. These were the trials of interest and made up 75% of the total number of trials. In order to ensure that the participants must attend to the contestant's belief state, regardless of whether it was true or false, anti strategy trials, termed herein 'fillers', formed 25% of the presented trials (see Hartwright et al., 2012 for further discussion). Here, the response probe consisted of a full colour photograph of the contestant, which had been blurred using a Gaussian smoothing kernel of 10 pixels FWHM. A black question mark obscured part of the contestant's face. Participants had been trained to indicate the true location of the prize when this format of response probe was shown. These fillers did not form any part of the analyses presented here.

Images of the contestants were taken from the Radboud Faces Database, (Langner et al., 2010). 28 contestant's featured in the experiment (all Caucasian; 14 male), where each face was shown on 6 occasions throughout the experiment, once per block. Each facial expression – happy, sad, neutral – was shown twice for each face. Each image consisted of a head and shoulders shot on a plain grey background. All contestants were wearing a plain black t-shirt. Each participant viewed a total of 168 rounds of the game show, made up of 126 trials of interest and 42 anti-strategy fillers, each with a unique prize, presented over the 6 blocks. Fig. 7 outlines the structure of the social judgements paradigm.

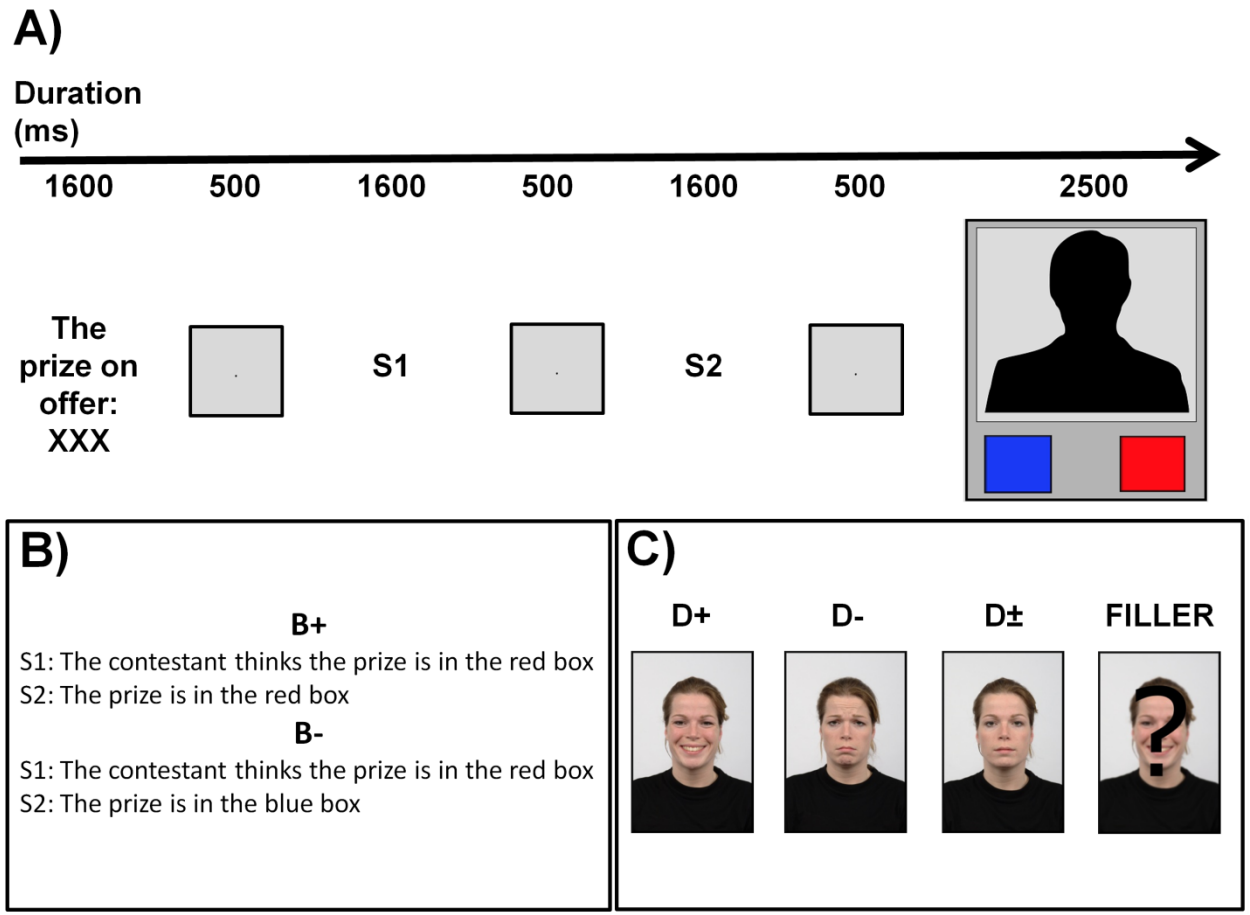

Fig. 7. Method: Social Judgements Paradigm
(Panel A) Schematic example of a single trial. The left/right presentation of the red/blue box was randomised. Where XXX is written for the prize on offer, this would name a unique item for each trial, e.g., The prize on offer: hot tub **(Panel B)** Example statements for true (B+) and false (B-) belief scenarios. The temporal order of these statements was randomised. **(Panel C)** From left to right, example response probe for approach- (D+), avoidance- (D-), unspecified-desire (D±) and filler trials.

Data Acquisition

Data were acquired in a single session using a 3T Philips Achieva scanner, with an 8 channel head coil. Whole brain coverage was achieved with the following parameters: TR = 2.5 s, TE = 35 ms, acquisition matrix = 96 x 96, flip angle = 83°, SENSE factor = 2. 232 T2*-weighted echo-planar imaging (EPI) volumes were obtained per block of the experiment, each of which consisting of 42 axial slices obtained consecutively in a bottom up sequence, reconstructed voxel size = $3x3x3mm^3$. Four dummy volumes were acquired at scan time; these were removed prior to image reconstruction. Following acquisition of the functional data, a T1-weighted anatomical image was acquired (3D TFE, sagittal orientation, TR=8.4 ms, TE=3.8, matrix size 288x288, 175 slices, reconstructed voxel size = $1\times1\times1mm^3$). During the acquisition of functional data, Presentation software (v. 14.1; Neurobehavioral Systems, CA) was used to display the stimuli and record the behavioural response data simultaneously.

Whole Brain Analysis

The FMRIB software library (FSL version v.5.98; FMRIB, Oxford, www.fmrib.ox.ac.uk/fsl) was used to perform all preprocessing and statistical analyses. Preprocessing of the functional data consisted of slice timing (regular up) and motion correction (MCFLIRT). High-pass filtering was conducted on the blood oxygen level dependent (BOLD) signals using a Gaussian weighted filter of 30 s. Spatial smoothing was then applied using a 5mm full-width-half-maximum kernel. The functional data were registered to their respective structural images and transformed to the Montreal Neurological Institute (MNI) reference brain using a 7-DoF linear transformation (FLIRT).

The modelling approach replicates the procedure outlined in Hartwright et al., (2012), which allows direct comparison following the minimal change to our previous paradigm. Six

explanatory variables (EVs) of interest – B+D+, B+D-, B+D±, B-D+, B-D-, B-D± – were modelled to reflect the six experimental conditions. The onset of each EV was time locked to the button response and reflected an arbitrary duration of 100ms. Due to anticipated differences in reaction times as a function of experimental condition, this approach ensured that activation reflected the decision making phase within the experimental sequence. This approach mirrors Hartwright et al. (2012) which was adopted following careful inspection of time series data. Each EV was convolved with a gamma derived haemodynamic response function (HRF) within a general linear model framework (GLM). The time series prior to the onset of the response probe was modelled as a regressor of no interest and orthogonalised with respect to the main EVs. Motion parameters and filler trials were also modelled as regressors of no interest. Higher level modelling was used to aggregate the data across participants within a mixed effects (ME) model using cluster based thresholding at voxel Z > 2.5, cluster $p_{corr} < 0.001$. Note that this particular threshold was applied for ease of comparison with the earlier published version of this paradigm. This final whole brain result reflected a 2x3 repeated measures ANOVA with Belief (B+/B-) and Desire (D+/D-/D±) as within subjects factors, plus 8 contrasts for directional tests comparing the levels of each main factor (e.g., B+ > B-, B- > B+ etc).

Contrast Masking Analysis

In order to demonstrate voxels which were preferentially active for each of the three levels within the factor of desire, using FSL's command line tools (fslmaths), the corrected, thresholded data from the directional whole brain analysis were used as inputs to generate 3 masks, $D+_{pref}$, $D-_{pref}$ and $D\pm_{pref}$. This was done by computing a logical AND which collapsed across the pairs of directional contrasts for each level (i.e., $D+_{pref}$ = D+ > D- AND

D+ > D±; $D-_{pref}$ = D- > D+ AND D- > D±; $D±_{pref}$ = D± > D+ AND D± > D-). Note that this analysis was not required for the factor of belief, as the directional contrasts serve this purpose ($B+_{pref}$ = B+ > B-; $B-_{pref}$ = B- > B+). The mean effect across all conditions was also computed for significantly active voxels within a bisected region of interest (slices X = -10 through to +10; MNI coordinates). This enabled identification of voxels which were preferentially active for unspecified-desire (D±) versus all other belief (B+/B-) and desire (D+/D-) conditions within mPFC.

RESULTS

Behavioural Data

All reaction times (RTs) were recorded from the onset of the response probe. Any incorrect responses were removed for RT analysis. Note that correct responses are only applicable in D+/D- trials as D± requires a subjective judgement. A 2x3 repeated measures ANOVA was conducted on the remaining RT data, with Belief (B+/B-) and Desire (D+/D-/D±) as within subjects factors. This revealed significant main effects of Belief, where B- > B+ ($F(1,19) = 166.65$, $p < 0.001$, $\eta^2 = 0.90$) and Desire, where D± > D- > D+ ($F(1,19) = 99.40$, $p < 0.001$, $\eta^2 = 0.84$), and a significant interaction ($F(2,38) = 4.86$, $p < 0.05$, $\eta^2 = 0.20$). Simple effects analyses revealed significant effects of belief at each level of the factor of the desire, and significant effects of desire at the two levels of the belief factor, (all *p*s < 0.01); however, with the interaction being accounted for by the effect of belief being largest when desires were negative. A further two-way ANOVA was computed on the error data, with Belief (B+/B-) and Desire (D+/D-) as repeated measures. This identified a main effect of Belief, where errors B- > B+ ($F(1,19) = 6.68$, $p < 0.05$, $\eta^2 = 0.26$) and Desire, where errors D-

> D+ $F(1,19) = 8.35$, $p < 0.01$, $\eta^2 = 0.31$). No interaction was identified ($F(1,19) = 0.20$, $p = 0.66$, $\eta^2 = 0.01$). Fig. 8 summarises the mean RT (Panel A) and accuracy data (Panel B) .

fMRI Data

Whole Brain Analysis

A 2x3 repeated measures ANOVA identified main effects of Belief (B+/B-) and Desire (D+/D-/D±) but no interaction between the two factors. Manipulation of an agent's belief state replicated our previously published findings (Hartwright et al., 2012), yielding regions regularly implicated in ToM such as bilateral TPJ and precuneus. Variation of an agent's belief state modulated considerable portions of the frontal cortex including bilateral dorsolateral and ventrolateral prefrontal cortices spanning middle frontal and inferior frontal gyri, extending to orbital frontal cortex. This factor also recruited bilateral dorsal medial frontal regions comprising superior frontal, dorsal anterior cingulate and dorsal paracingulate gyri (Table 4; red shading Fig. 9). Similar to belief reasoning, manipulation of an agent's desire state also recruited bilateral temporoparietal junction. However, lateral and medial prefrontal regions were recruited more extensively; thus, encompassed bilateral frontal poles on the lateral and medial surface, as well as rostral medial frontal regions including more ventral sections of the anterior cingulate and paracingulate gyri. Unlike the belief condition, variation of an agent's desire state also saw extensive recruitment of occipital regions spanning bilateral occipital poles to anterior occipital regions such as the calcarine cortex (Table 4; green shading Fig. 9).

(A) **(B)**

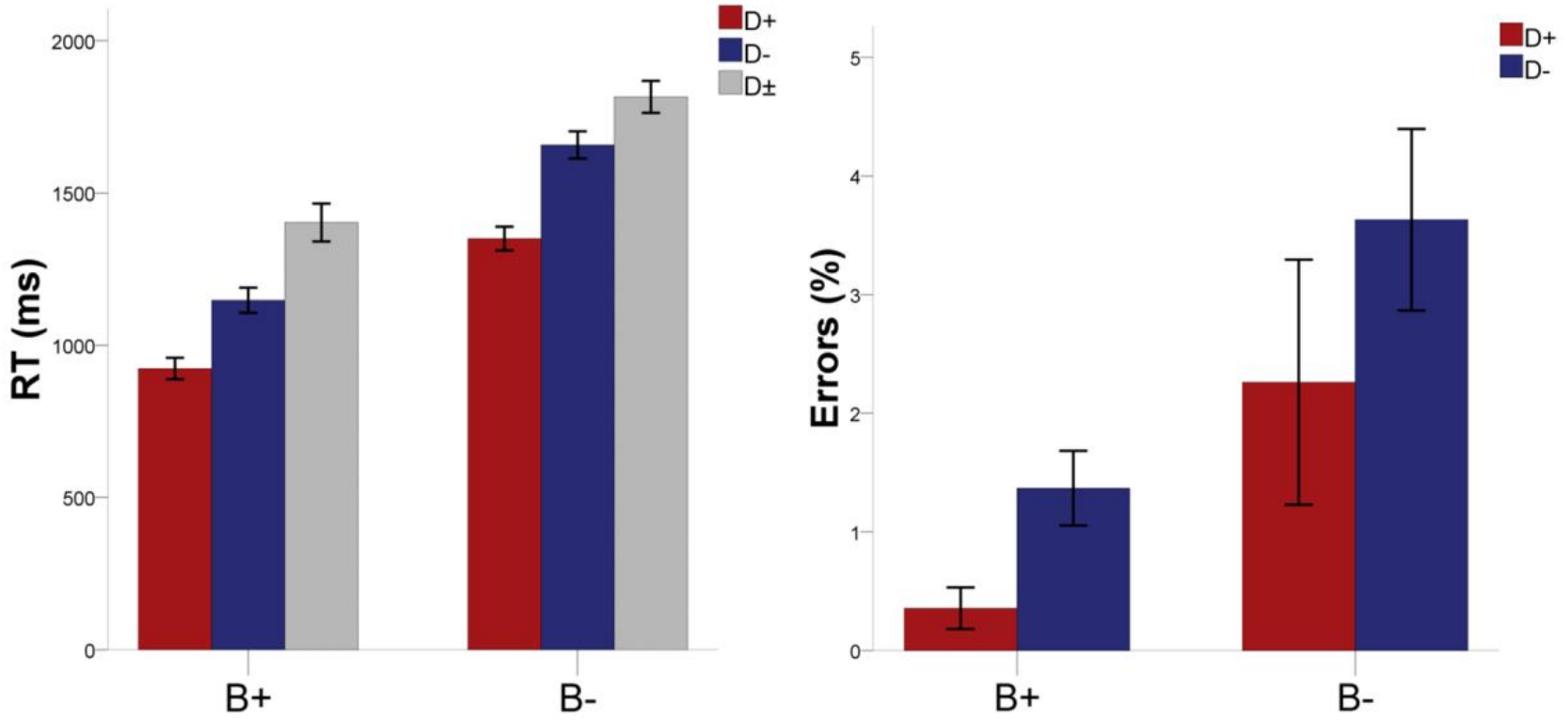

Fig. 8. Results: Social Judgments Task Behavioural Data
Error bars reflect +/-1 SE of the mean. **(Panel A)** Group mean reaction time per condition for correct responses (ms). **(Panel B)** Group mean percentage of errors; error data not applicable to D±.

Table 4 Factorial Analysis of Belief and Desire

Region	Hemi	Brodmann Area	MNI coordinates			Z-value
			x	y	z	
Main effect of Belief						
Temporoparietal Junction	R	22	54	-56	26	6.47
Precuneus Cortex	R	7	2	-66	48	5.76
Orbital Frontal Cortex	L	47	-32	26	-2	5.42
Temporoparietal Junction	L	40	-52	-52	32	5.41
Insular Cortex	R	47	46	16	-6	5.22
Middle Frontal Gyrus	R	44	50	20	38	5.18
Paracingulate Gyrus	L	8	-4	20	48	5.09
Frontal Pole	R	46	38	52	18	5.03
Inferior Frontal Gyrus, pars opercularis	L	45	-48	16	0	4.97
Precuneus Cortex	L	7	-6	-66	54	4.94
Paracingulate Gyrus	R	32	2	42	28	4.86
Middle Frontal Gyrus	L	44	-46	14	36	4.80
Insular Cortex	L	47	-40	16	-6	4.40
Superior Frontal Gyrus	R	9	2	40	42	4.38
Inferior Frontal Gyrus, pars opercularis	R	48	52	18	4	4.34
Supramarginal Gyrus, anterior division	L	40	-54	-40	38	4.21
Lateral Occipital Cortex, superior division	R	39	44	-60	52	4.18
Supramarginal Gyrus, posterior division	R	40	48	-46	42	4.12
Postcentral Gyrus	L/R	5	0	-54	72	3.52
Superior Frontal Gyrus	L	6	-2	14	68	3.29
Cingulate Gyrus, anterior division	L	32	-8	40	16	3.16
Main effect of Desire						
Paracingulate Gyrus	R	8	2	24	48	8.34
Paracingulate Gyrus	L/R	8	0	28	42	8.15
Occipital Pole	R	18	22	-98	-2	7.42
Occipital Pole	L	18	-24	-94	-8	7.16
Superior Frontal Gyrus	L	8	-8	30	58	6.34

Orbital Frontal Cortex	R	47	36	24	-6	5.78
Occipital Fusiform Gyrus	R	18	18	-84	-8	5.66
Intracalcarine Cortex	R	17	14	-84	2	5.63
Superior Frontal Gyrus	R	8	4	54	40	5.61
Occipital Fusiform Gyrus	L	18	-14	-84	-12	5.59
Middle Frontal Gyrus	R	45	52	28	24	5.57
Lateral Occipital Cortex, superior division	L	19	-22	-86	20	5.56
Inferior Frontal Gyrus, pars opercularis	R	48	54	20	6	5.50
Orbital Frontal Cortex	L	47	-38	22	-8	5.29
Frontal Pole	R	46	24	56	22	5.07
Cingulate Gyrus, anterior division	L/R	32	0	44	14	5.01
Temporoparietal Junction	L	39	-46	-58	44	4.70
Inferior Frontal Gyrus, pars triangularis	L	45	-50	20	0	4.43
Temporal Pole	L	38	-48	16	-10	4.28
Inferior Frontal Gyrus, pars triangularis	R	45	54	34	12	4.19
Supramarginal Gyrus, posterior division	R	40	52	-44	48	4.04
Frontal Pole	L	47	-50	34	-20	4.03
Temporoparietal Junction	R	22	60	-58	26	3.96
Middle Frontal Gyrus	L	44	-48	16	36	3.92
Supramarginal Gyrus, anterior division	L	40	-42	-38	38	3.88
Supramarginal Gyrus, posterior division	L	40	-42	-44	38	3.88
Inferior Frontal Gyrus, pars opercularis	L	45	-52	22	22	3.69
Temporal Pole	L	38	-40	28	-24	3.62
Lateral Occipital Cortex, superior division	R	39	54	-62	34	3.54
Supramarginal Gyrus, anterior division	R	2	54	-32	48	3.43
Postcentral Gyrus	L	40	-32	-36	42	2.59

Note. Regions identified using F-contrasts in a 2-way repeated measures ANOVA with factors of Belief (B+/B-) and Desire (D+/D-/D±). Table lists local maxima for cortical regions which are modulated by varying belief status (true/false) and desire status (approach/avoid/unspecified) Z > 2.5, pcorr < 0.001. All anatomically unique local maxima (with minimum peak separation of 5mm) are listed. Brodmann Areas are approximate.

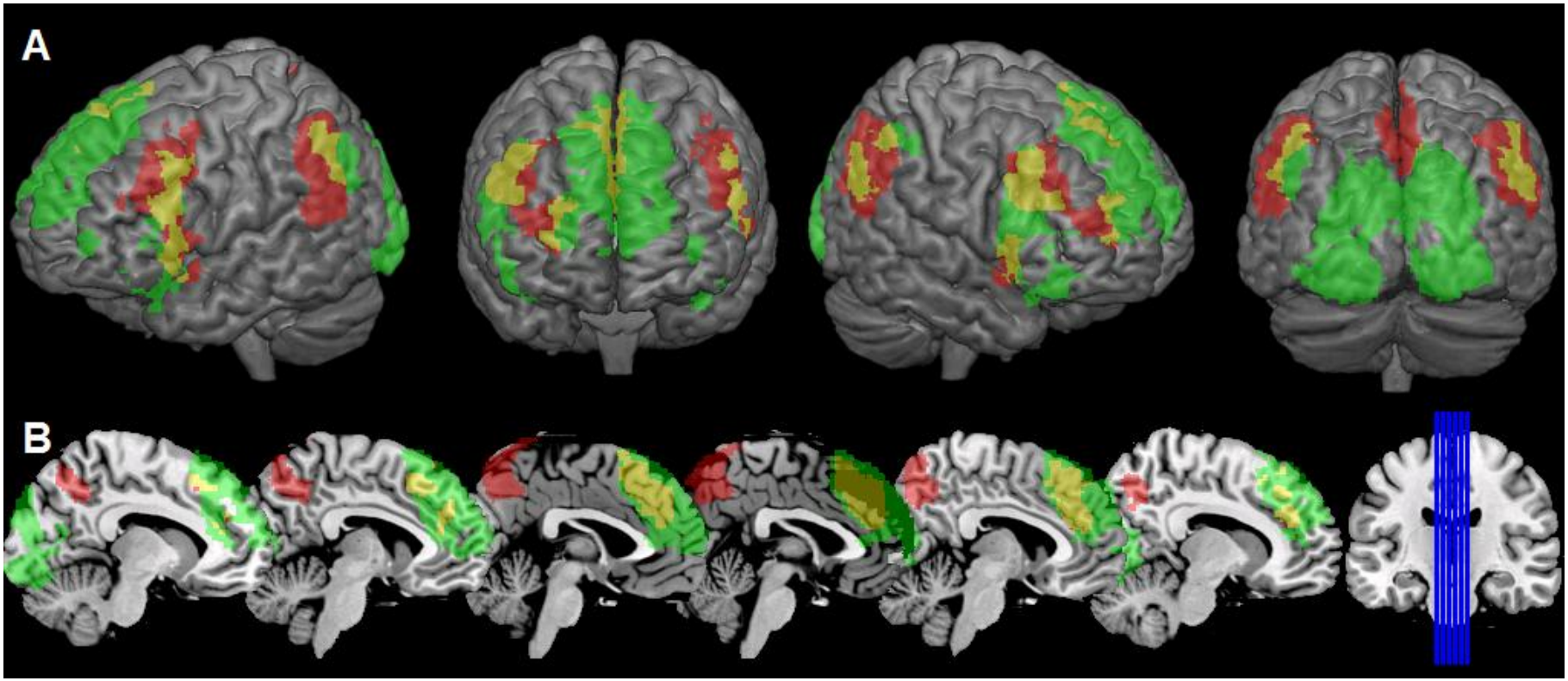

Fig. 9. Results: Social Judgments Task Whole Brain Analysis
Result from 2x3 repeated measures ANOVA whole brain analysis with Belief (B+/B-; red) and Desire (D+/D±/D-; green) as within-subjects factors. Yellow areas indicate regions recruited by both factors (B/D). The group data are overlaid on the MNI brain template, showing significantly activated voxels where Z > 2.5, $p_{corr} < 0.001$. Maps reflect Z-corrected F-stat images and are displayed in neurological convention, where left is represented on the left side of the image. **(Panel A)** Activation maps highlighting modulation on the lateral surface. Images from left to right show left, anterior, right, and posterior views of the cortex respectively. **(Panel B)** Selected slices highlight modulation in medial frontal regions. Slices from left to right, x = -10, -6, -2, 2, 6, 10.

Directional and Contrast Masking Analysis

A series of directional contrasts (Table 5) demonstrated that bilateral temporoparietal junction, superior parietal and occipital cortices, plus lateral and dorsal medial frontal regions, were typically more responsive when applying false- over true-belief reasoning to an agent (Fig. 20A, $B-_{pref}$). The only regions that were preferentially active for true- over false-belief reasoning were the left occipital pole and occipital cortex (Fig. 20A, $B+_{pref}$). For desire based reasoning, contrast mask analyses indicated that bilateral occipital cortices and bilateral pre and post central gyri were preferentially responsive when applying an approach- versus avoidance- or unspecified-desire (Fig. 20B, $D+_{pref}$). When the agent expressed an avoidance- versus an approach- or unspecified-desire, right precuneus was the only region to be preferentially recruited (Fig. 20B, $D-_{pref}$). A large area covering medial and lateral prefrontal cortex was highlighted to be most responsive when the agent's desire was unspecified, versus to approach or avoid. Medial frontal activation spanned anterior cingulate, dorsal and rostral medial prefrontal cortices, extending laterally to bilateral frontal poles (Fig. 20B, $D\pm_{pref}$). As shown in Fig. 20C, rostral medial prefrontal cortex was preferentially active for unspecified desire over and above all of the other belief and desire states.

Table 5 Directional Contrasts Within the Factors of Belief and Desire

Region	Hemi	Brodmann Area	MNI coordinates			Z-value
			x	y	z	
Belief						
B+ > B-						
Occipital Pole	L	18	-20	-94	-10	5.52
Lateral Occipital Cortex, superior division	L	18	-24	-88	18	5.31
Lateral Occipital Cortex, inferior division	L	19	-36	-90	-12	4.02
B- > B+						
Temporoparietal Junction	R	22	50	-50	26	7.18
Precuneus Cortex	R	7	2	-66	48	5.87
Temporoparietal Junction	L	40	-52	-52	30	5.76
Insular Cortex	R	47	44	16	-6	5.63
Frontal Orbital Cortex	L	47	-32	26	-2	5.54
Supramarginal Gyrus, posterior division	L	40	-44	-50	42	5.43
Precuneus Cortex	L	7	-8	-64	50	5.38
Occipital Pole	L/R	17	0	-92	-12	5.36
Middle Frontal Gyrus	R	44	50	20	38	5.31
Paracingulate Gyrus	L	8	-4	20	48	5.21
Inferior Frontal Gyrus, pars opercularis	L	48	-54	16	0	5.20
Frontal Pole	R	46	38	52	18	5.17
Middle Frontal Gyrus	L	9	-50	14	44	5.09
Paracingulate Gyrus	R	32	2	42	28	5.00
Inferior Frontal Gyrus, pars triangularis	L	45	-60	22	8	4.96
Temporal Pole	L	38	-52	16	-10	4.83
Frontal Operculum Cortex	L	47	-44	18	-4	4.71
Superior Frontal Gyrus	L/R	8	0	22	56	4.57
Superior Frontal Gyrus	R	9	2	40	42	4.53
Inferior Frontal Gyrus, pars opercularis	R	48	52	18	4	4.49
Lateral Occipital Cortex, superior division	R	39	44	-58	40	4.38
Supramarginal Gyrus, anterior division	L	40	-54	-40	38	4.37
Lingual Gyrus	R	18	4	-84	-16	3.97
Postcentral Gyrus	L/R	5	0	-54	72	3.70

Desire

D+ > D-

Occipital Pole	L	18	-22	-94	-8	8.00
Occipital Pole	R	18	22	-96	-2	7.49
Lateral Occipital Cortex, inferior division	L	19	-36	-90	-12	5.60
Lateral Occipital Cortex, superior division	R	19	12	-86	44	5.24
Precentral Gyrus	L	6	-54	-2	46	4.85
Precentral Gyrus	R	6	50	-8	54	4.48
Postcentral Gyrus	R	4	4	-36	56	4.26
Superior Parietal Cortex	R	5	20	-50	66	4.17
Postcentral Gyrus	L	2	-28	-40	66	4.09

D+ > D±

Occipital Pole	L	18	-20	-96	-6	7.96
Occipital Pole	R	18	22	-98	0	7.84
Occipital Fusiform Gyrus	R	18	20	-82	-10	6.38
Lingual Gyrus	R	18	16	-88	-6	6.11
Occipital Fusiform Gyrus	L	18	-14	-84	-12	6.01
Intracalcarine Cortex	R	17	14	-84	2	6.00
Precentral Gyrus	L	6	-54	-2	46	5.27
Precentral Gyrus	R	4	52	-4	38	4.61
Postcentral Gyrus	R	3	24	-38	76	4.22
Postcentral Gyrus	L	3	-24	-40	70	4.17
Superior Parietal Cortex	R	5	20	-50	68	4.09
Superior Frontal Gyrus	R	6	16	2	72	3.94
Superior Temporal Gyrus, anterior division	L	21	-56	2	-12	3.53
Planum Polare	L	38	-58	2	-2	3.52
Central Opercular Cortex	L	48	-50	-2	8	3.25

D- > D+

Paracingulate Gyrus	R	8	2	24	48	8.06
Superior Frontal Gyrus	R	9	2	40	42	7.17
Superior Frontal Gyrus	L	8	-8	30	46	6.62
Frontal Orbital Cortex	R	47	34	22	-8	6.50
Middle Frontal Gyrus	R	45	52	28	24	6.48
Insular Cortex	R	47	34	24	0	6.45
Temporoparietal Junction	R	22	52	-56	26	6.39
Inferior Frontal Gyrus, pars opercularis	R	48	54	20	6	6.37
Frontal Orbital Cortex	L	47	-32	24	-8	6.08

Temporoparietal Junction	L	40	-48	-52	42	5.61
Precuneus Cortex	R	7	4	-68	42	5.35
Supramarginal Gyrus, posterior division	R	40	52	-46	48	5.34
Supramarginal Gyrus, posterior division	L	40	-46	-44	44	5.27
Precuneus Cortex	L/R	7	0	-68	54	4.65
Lateral Occipital Cortex, superior division	R	39	44	-58	50	4.53
Precuneus Cortex	L	7	-8	-64	50	4.15
Lateral Occipital Cortex, superior division	L	39	-50	-68	44	2.72
D- > D±						
Lateral Occipital Cortex, superior division	R	19	28	-80	24	4.43
Supramarginal Gyrus, posterior division	L	22	-56	-46	10	4.42
Occipital Pole	R	18	14	-88	22	4.39
Supramarginal Gyrus, anterior division	L	40	-58	-38	32	4.36
Intracalcarine Cortex	R	17	10	-70	14	4.29
Occipital Pole	L	18	-18	-96	-2	4.14
Cuneal Cortex	R	18	4	-84	32	4.11
Precuneus Cortex	L	7	-4	-58	56	3.94
Superior Temporal Gyrus, posterior division	L	21	-54	-30	-2	3.89
Middle Temporal Gyrus, posterior division	L	21	-58	-24	-8	3.85
Lateral Occipital Cortex, inferior division	L	37	-46	-70	2	3.73
D± > D+						
Paracingulate Gyrus	R	8	2	24	48	8.67
Paracingulate Gyrus	L/R	8	0	28	42	8.48
Paracingulate Gyrus	L	32	-4	34	36	8.31
Superior Frontal Gyrus	R	8	4	42	50	7.37
Superior Frontal Gyrus	L	8	-4	28	60	7.11
Insular Cortex	R	47	36	22	-6	6.17
Lateral Occipital Cortex, superior division	L	39	-46	-60	36	5.63
Temporoparietal Junction	L	39	-46	-58	42	5.29
Cingulate Gyrus, posterior division	L	23	-2	-54	24	4.51
Lateral Occipital Cortex, superior division	R	22	60	-60	26	4.48
Temporoparietal Junction	R	39	50	-58	28	4.37
Supramarginal Gyrus, posterior division	L	40	-46	-50	52	3.83
Supramarginal Gyrus, anterior division	R	2	54	-28	44	3.73
Supramarginal Gyrus, anterior division	L	2	-46	-38	44	3.72
Precuneus Cortex	R	7	2	-66	34	3.64
Supramarginal Gyrus, posterior division	R	40	50	-44	54	3.52

Cingulate Gyrus, posterior division	L/R	23	0	-16	28	3.48
Cingulate Gyrus, anterior division	L/R	23	0	-10	28	3.21
Cingulate Gyrus, posterior division	L	29	-4	-48	14	2.85
D± > D-						
Paracingulate Gyrus	R	32	2	50	26	5.97
Paracingulate Gyrus	L/R	32	0	38	34	5.97
Frontal Pole	L/R	10	0	60	30	5.72
Frontal Pole	L	9	-16	40	48	5.39
Superior Frontal Gyrus	L	8	-2	38	50	5.08

Note. Table lists local maxima for cortical regions identified using a series of directional t-contrasts, where Z > 2.5, $p_{corr} < 0.001$. All anatomically unique local maxima (with minimum peak separation of 5mm) are listed. Brodmann Areas are approximate.

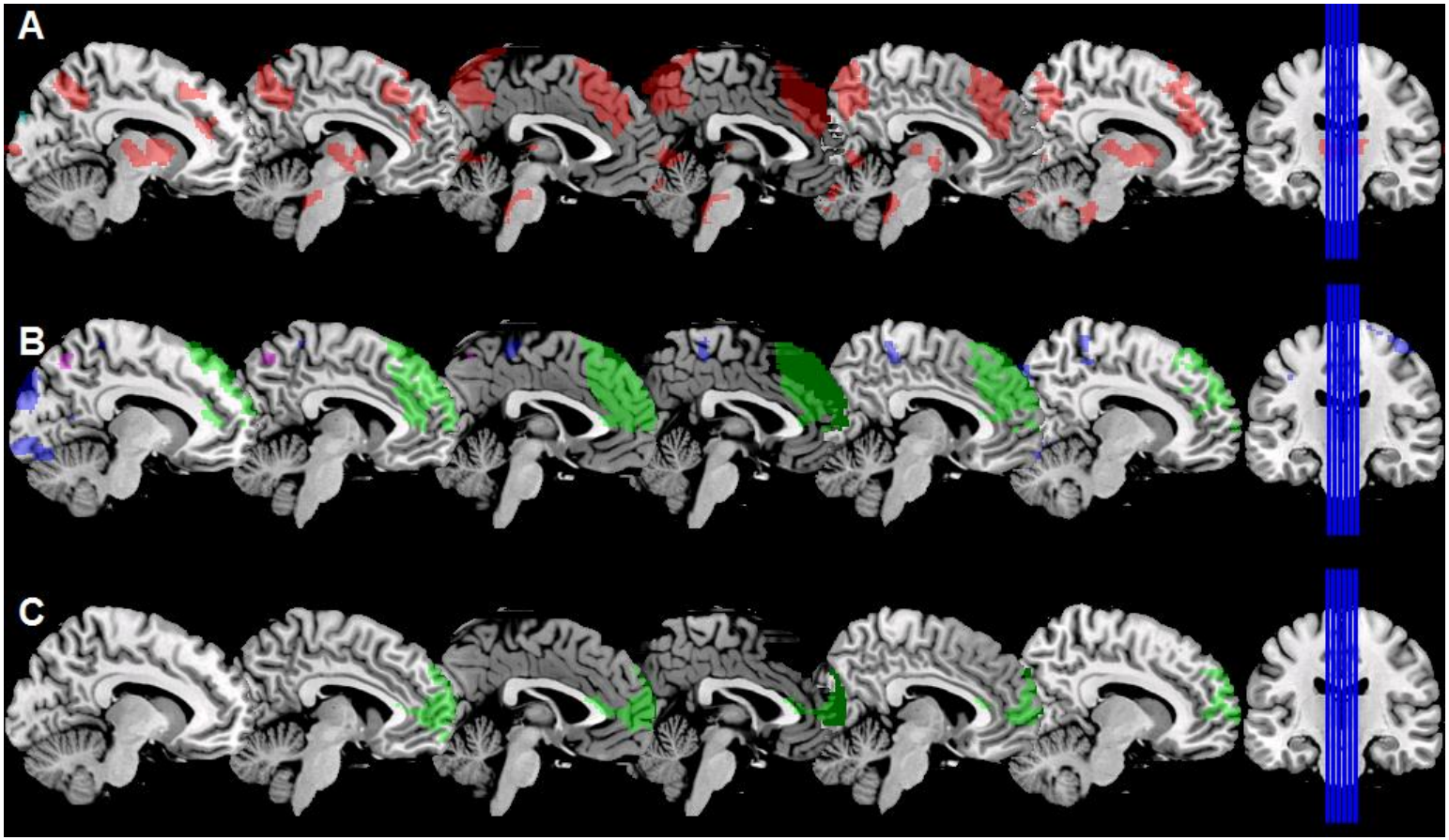

Fig. 20. Results: Social Judgments Task Contrast Masking Analysis
The group data are overlaid on the MNI brain template, showing significantly activated voxels where Z > 2.5, $p_{corr} < 0.001$. Slices from left to right, x = -10, -6, -2, 2, 6, 10 respectively. Maps reflect Z-corrected t-stat images. **(Panel A).** Voxels which are preferentially active during true- versus false-belief reasoning ($B+_{pref}$; cyan); false- versus true-belief reasoning ($B-_{pref}$; red). **(Panel B).** Voxels which are preferentially active during approach- versus unspecified- AND avoidance-desire ($D+_{pref}$; blue); avoidance- versus approach- AND unspecified-desire ($D-_{pref}$; magenta); unspecified- versus approach- AND avoidance-desire ($D\pm_{pref}$; green). **(Panel C)** Voxels within the medial frontal cortex which are preferentially active for unspecified-desire versus all other belief and desire conditions.

DISCUSSION

The diverse array of social and non-social tasks that activate mPFC has meant that the precise role of this region in ToM has remained vague (Rothmayr et al., 2011). We employed an analysis of common features of ToM tasks to distinguish roles that mPFC might serve for *representation, control* and *reasoning*. The need to represent mental states was present in all task conditions, while the task made it possible for the first time to manipulate demands on control and reasoning within a single study. Our results suggest that dorsal and rostral regions of mPFC play distinctive roles in ToM control and ToM reasoning respectively, and that these patterns are consistent with the proposed functions of these regions in non-social tasks.

Conflict Monitoring, Control and the dmPFC

On the basis of previous behavioural and neuroimaging work, we expected that greater control would be required when predicting action based on false- versus a true-belief, or a desire to avoid versus approach an object. Behavioural data from the current study were consistent with these predictions. The neuroimaging results converge with the general executive literature in pinpointing dmPFC, comprising dACC and paracingulate gyrus (PCG), in supporting these more cognitively effortful scenarios (Botvinick et al., 2004; Botvinick et al., 1999; Bush et al., 2000). Factorial analysis (Table 4, Fig. 9) showed that dmPFC was modulated by manipulating the content of specific ToM states. Investigation of the directional contrasts (Table 5) highlighted that these main effects were driven by those mental state concepts where the greatest need for control existed, such as false belief, avoidance- and unspecified-desire. Notably, the novel, unspecified desire condition attracted the greatest increase in response latencies, and made greater demands on dmPFC than both avoidance-

and approach-desire reasoning. We propose that this result is consistent with our suggestion that dmPFC serves conflict detection in support of control processes, because in order to predict the behaviour of the agent with an unspecified desire, participants would have to withhold any response until they had determined what they thought the agent's preferences might be. Here, then, conflict exists not only between competing outcomes, such as the undesirable- versus the desirable-outcome, but also potentially between what the participant would do, and what someone like the agent would do in that particular situation. Taken together, then, these data are further evidence that dmPFC serves a very general control function, with more specific functions – such as inhibition of self-perspective – supported by other neural regions.

ToM Reasoning and the rmPFC

Also of interest was the role of rmPFC in ToM. Existing literature, together with the task analysis presented here, suggests two possible roles for this region, and our task was designed to distinguish between them. First, the consistency with which rmPFC is recruited for ToM in previous research has led some authors to suggest that this region serves the function of *representing* mental states (Amodio & Frith, 2006; Frith & Frith, 2006; Frith & Frith, 2003). All conditions of our current paradigm required representation of the character's mental states, and so this interpretation of the role of rmPFC does not predict any variation in activation across conditions. Second, a growing literature indicates that thinking beyond the stimuli presented recruits rmPFC even in non-social contexts (Gilbert et al., 2007; Jenkins & Mitchell, 2009). Thinking beyond the stimuli, and in particular so-called "abductive" inference to the best explanation, is a frequent requirement of ToM, both in tasks and outside of the laboratory. However, it is not a necessary feature, and it was not present in the belief

factor of the current task, while the desire factor included one level that required abductive reasoning (D±) and two levels that only required deductive reasoning (D- and D+).

Consistent with Hartwright et al. (2012), factorial analysis identified that manipulating an agent's belief state did not modulate rmPFC (Table 4, Fig. 9). Thus, there was no difference in how reasoning deductively about an agent with a true- or false-belief state was handled by this region. In contrast to this, manipulation of an agent's desire state was shown to modulate rmPFC. Note that this was not the case in our previous study, which did not require abductive reasoning in any condition. Directional and contrast masking analyses (Table 5, Fig. 20) were used to clarify which of the variations in mentalizing was driving this effect. rmPFC was shown to respond preferentially when reasoning about an agent whose desire was unspecified (D±), over and above any of the other deductive belief and desire conditions (Fig. 20C). Collectively, these data suggest that rmPFC is responsive to the requirement to reason abductively about mental states.

These findings converge with Jenkins and Mitchell (2009), who found that comprehension of a story whose causal structure was ambiguous or incomplete, rather than unambiguous and complete, preferentially recruited mPFC, including rmPFC. Such effects were found irrespective of whether the stories required inferences about a character's mental states, and indeed it is unclear in this study whether rmPFC was recruited for the ToM inferences themselves or just for general comprehension of an ambiguous context. The current study provides important clarity on this point, by showing that rmPFC is indeed recruited for ToM inferences specifically in cases where abductive rather than deductive reasoning is required. In the broader social context, Van Overwalle (2009) notes that studies which invite richer inferences, such as trait ascription, recruit mPFC. Relatedly, Quadflieg et al. (2009) demonstrated that rmPFC is recruited when reasoning about the type of person

(male/female/either), versus the type of place (indoors/outdoors/either) that is likely to be associated with an activity, such as mowing the lawn or watching talk shows. Thus, rmPFC was seen as an important neural substrate of the access and assignation of stereotype information. It is important to highlight, however, that this does not conflict with our assertion that rmPFC supports a general process that is engaged when reasoning abductively. A considerable literature demonstrates the automaticity of trait inferences and social categorisation (Greenwald & Banaji, 1995), for example, on the basis of an image of a face (Todorov, Said, Engell, & Oosterhof, 2008), or when primed subconsciously (Bargh, Chen, & Burrows, 1996). As such, all of our desire conditions featured the photographs of faces taken from a single database; our analyses would, therefore, subtract out those neural regions required for the attribution of stereotype schemas, as the potential for spontaneous trait ascription, including the automatic generation of stereotypes, is constant across all conditions. Our unspecified desire condition, on the other hand, is the only condition to require an abductive inference on the basis of such ascriptions. When considered alongside a literature which implicates rmPFC in autobiographical thinking, for example, in terms of imagining past or future events versus simply recalling such occurrences, prospection and the default mode network (see reviews by Schacter et al., 2012; Spreng et al., 2009), the commonality across these, and Jenkins and Mitchell (2009), is a shared process which reflects the assignation of information which is obtained through a rich, inferential process. The present paradigm varied the requirement for this process, by including a single, abductive reasoning condition alongside a series of deductive reasoning conditions.

Cognitive versus Affective ToM.

Qualitative reviews of the literature suggests a functional subdivision within mPFC, where a dorsal/rostral boundary may delineate cognitive- versus affective-ToM respectively (Abu-Akel & Shamay-Tsoory, 2011; Amodio & Frith, 2006; Carrington & Bailey, 2009; Lieberman, 2007). Thus, belief reasoning would be expected to recruit more dorsal regions of mPFC, whereas desire reasoning would recruit rostral regions. Whilst the current data might initially appear to favour this distinction, we suggest that a simple cognitive/affective division provides less explanatory power for our data than the task analysis proposed here.

First, the present study suggests that it is likely to be the processing requirements within particular ToM concepts that modulate dmPFC (e.g., true versus false belief), rather than the cognitive or affective nature of the ToM concept. Our data identify that cognitively effortful situations involving false belief, avoidance- or unspecified-desire reasoning make greater demands on dmPFC than less effortful ToM situations such as true belief or approach desire. We suggest that this effort, seen in increased response latencies and errors, is a reflection of increased conflict between alternative predictions for the agent. Thus, increased effort is associated with increased demand on dmPFC, regardless of the type of mental state being represented.

Second, whilst only our affective (desire) condition recruited rmPFC, this region was preferentially engaged as a function of the reasoning demands within this condition, rather than the mere requirement to infer desires. Specifically, a context that required abductive inference about desire was associated with increased demand on rmPFC, compared with conditions that only required deductive inferences about desire. Our data show that rmPFC is brought in to serve context specific *reasoning* processes, such as when mentalizing beyond the information presented is required.

Representing Mental States

By definition, ToM requires people to hold in mind representations of mental states, and questions about this representational aspect of ToM have dominated thinking in the developmental, cognitive, comparative and neuroscience literatures (Call & Tomasello, 2008; Fodor, 1992; Leslie, 1987; Saxe & Powell, 2006). However, identification of the neural basis of such representations has proved a surprisingly elusive target, with ongoing debates about the relative specificity of mPFC versus TPJ for such representations (Aichhorn et al., 2009; Amodio & Frith, 2006; Saxe & Kanwisher, 2003; Saxe & Powell, 2006; Saxe & Wexler, 2005; Scholz et al., 2009). The present study was not designed as a strong test of the neural correlates of representing mental states, as we did not include conditions without mental states for comparison. However, the current findings do add to a growing body of evidence suggesting that the mere representation of mental states is only part of the neurocognitive basis of ToM, in two important ways. First, other functional processes for cognitive control and reasoning are integral to ToM, and recruit neural regions supporting these processes in ways that can be predicted from functional analysis of ToM tasks. Second, even if a consensus does emerge on neural regions that are involved in representing mental states, it seems unlikely that this function will be sufficient to explain patterns of activity in those neural regions during ToM tasks. In the present study, mental states needed to be represented in all conditions, and yet we observed condition-wise variation in activity in the neural regions most often suggested to be the neural basis of representing mental states (rmPFC, and bilateral TPJ). Such variation can be understood by appeal to other functional aspects of ToM, such as the need for cognitive control, and the need for different kinds of reasoning. We suggest that this makes vivid the suggestion that ToM is subserved by a network, which may

be comprised of distinct functional and anatomical components, but whose activity can only be understood by considering the network as a whole, and the tasks in which it is engaged.

CHAPTER 4:

A CAUSAL ROLE FOR RIGHT VENTROLATERAL PREFRONTAL CORTEX IN SELF PERSPECTIVE INHIBITION? A PERTURBATION STUDY OF BELIEF-DESIRE REASONING

ABSTRACT

Converging lesion and neuroimaging evidence suggests that ventrolateral prefrontal cortex (vlPFC) is involved in inhibiting self perspective when making certain Theory of Mind (ToM) judgments, such as when the belief information conflicts with own knowledge, typically as in false belief reasoning. The present study used a continuous Theta Burst Stimulation (cTBS) protocol to depress cortical excitability in right vlPFC, in order to identify a causal role for this region in inhibition of self perspective. An interaction effect between stimulation site and belief state was expected, where a behavioural cost was anticipated in false belief reasoning following cTBS to vlPFC, versus a control site. Despite replicating behavioural effects consistent with prior work, no effect of cTBS was identified. The null result is discussed in the context of localization difficulties with applying cTBS to vlPFC, and new evidence which calls into question the reliability of this particular Transcranial Magnetic Stimulation (TMS) protocol.

INTRODUCTION

The neural basis of mental state attribution, or Theory of Mind (ToM), has been studied extensively with fMRI, with considerable focus on a fronto-parietal network comprising medial prefrontal cortex (mPFC) and temporoparietal junction (TPJ) (Carrington & Bailey, 2009; Lieberman, 2007; Mar, 2011). Activation of these regions is posited to reflect the basic neural mechanisms required for representing the mind state of an agent.

Perturbation techniques, such as TMS, provide the opportunity to directly identify a causal relationship between brain and behaviour in a specified region of interest, for example, by causing a 'virtual lesion' alongside its associated behavioural consequences (Pascual-

Leone et al., 2000). Within the wider social cognitive literature, TMS has been used to demonstrate a causal role for the inferior parietal lobule in facial recognition (Uddin, Molnar-Szakacs, Zaidel, & Iacoboni, 2006) and movement attribution (Preston & Newport, 2008), for example, where disruption to this region using repetitive (r)TMS impairs the ability to distinguish between self and other. Single pulse TMS to the primary motor cortex, M1, has evidenced that motor evoked potentials during action-observation are relative to the muscle group and force requirements of the action observed, thus adding credence to the idea of a human Mirror Neuron System (Alaerts et al., 2010).

Most brain stimulation studies specific to mentalizing have focussed on dorsolateral prefrontal cortex (dlPFC), mPFC and TPJ (see Hétu, Taschereau-Dumouchel, & Jackson, 2012). For example, rTMS to right dlPFC has been shown to reduce response latencies when making a belief inference, without any cost to response accuracy (Kalbe et al., 2010). rTMS to mPFC leads to faster recognition of emotions (Balconi & Canavesio, 2013) and an increase in the ability to make affective ToM judgements in individuals who report having low empathy, whilst impairing affective judgments in those who are highly empathic (Krause, Enticott, Zangen, & Fitzgerald, 2012). When applied to TPJ, rTMS impairs moral judgement, where knowingly causing harm to another is viewed as more permissible following stimulation. This effect is postulated to reflect interference in the ability to assimilate an agent's belief state in relation to the actual outcome of an intentionally, but failed, harmful act (Young, Camprodon, Hauser, Pascual-Leone, & Saxe, 2010).

As researchers begin to examine the composite processes of ToM, interest is moving towards those coactive, but lesser studied, neural regions and the role that they might play in a functioning ToM. Activation in vlPFC, for instance, is readily present across multiple neuroimaging studies of ToM (e.g., see reviews Mar, 2011; Spreng et al., 2009), yet little

examined. Nonetheless, it has been suggested that vlPFC reflects ToM situations where there is a need to inhibit one's own knowledge or experience (Hartwright et al., 2012; Lieberman, 2007; Samson et al., 2005; van der Meer et al., 2011). For example, using a series of ToM vignettes, Vogeley et al. (2001) were able to modulate right vlPFC by varying the presence of the experimental participant as a key character in each vignette, thus suggesting that this region is recruited as a function of self perspective. Following on from this, patient WBA, despite a sizeable right frontal lesion, was able to pass false belief tasks provided that the salience of his own perspective was minimised. Thus, when the prepotent, 'true' state of affairs was undisclosed, patient WBA could successfully demonstrate an understanding of someone else's belief state (Samson et al., 2005). These results were corroborated by van der Meer et al. (2011), who adapted the Samson et al. (2005) paradigm for use with neurologically intact participants. Their study identified bilateral vlPFC for high versus low salience scenarios and, within the same subjects, overlapping activation from a motor response inhibition task, suggesting some shared process between high salience ToM and response inhibition. Likewise, by contrasting classical unexpected transfer false belief scenarios with true belief scenarios, Hartwright et al. (2012) demonstrated that difficulty with certain mental states may, in part, reflect interference from incongruent self versus other knowledge, which is resolved in vlPFC. Using a single repeated measures ToM task, they highlighted that similarly difficult ToM states that do not feature incongruence between self and other perspectives do not modulate this region, suggesting that activation in vlPFC does not simply index the general difficulty of the task.

In sum, in terms of understanding the neural basis of self perspective inhibition in ToM, converging evidence from neuroimaging and neuropsychology implicate vlPFC in supporting this process. However, a causal role for vlPFC in self perspective inhibition is

difficult to determine given the form of data available to date. The very nature of neuroimaging is that any neural activation deemed 'significant' reflects a model of best fit between a neural signal – typically the BOLD response – and stimulation. Although, in principle, more powerful inferences can be drawn from lesion studies, working with human subjects prevents access to small, anatomically circumscribed lesions. Thus, researchers are typically working with individuals who have suffered damage to relatively large, heterogeneous areas of cortex, as was the case in Samson et al. (2005). TMS, therefore, provides the next logical step in identifying a causal role for vlPFC in self perspective inhibition.

Whilst TMS has been applied to vlPFC frequently in the study of language processes (Gough, Nobre, & Devlin, 2005; Nixon, Lazarova, Hodinott-Hill, Gough, & Passingham, 2004; Watkins & Paus, 2004) and executive functions, such as memory (Feredoes, Tononi, & Postle, 2006; Hong, Lee, Kim, Kim, & Nam, 2000) and response inhibition (Chambers et al., 2007; Verbruggen, Aron, Stevens, & Chambers, 2010), to our knowledge, it has not been used in this region in the context of self perspective inhibition. The present study used cTBS (Huang, Edwards, Rounis, Bhatia, & Rothwell, 2005), an offline TMS protocol which has been demonstrated effective for use on inferior frontal regions (Verbruggen et al., 2010). The cTBS protocol was selected as the neural time course of self perspective inhibition is little understood, therefore negating the use of online, single pulse TMS. Furthermore, the behavioural effects following cTBS, in particular, have been shown to last up to one hour, providing a longer lasting effect compared with other TMS protocols (Huang et al., 2005). On the basis of neuropsychological evidence suggesting that the right frontal cortex is involved in inhibiting self perspective (Samson et al., 2005), and that vlPFC, specifically, is associated with making ToM judgments when mental state information is distinct or incongruent with

own knowledge (Hartwright et al., 2012; van der Meer et al., 2011; Vogeley et al., 2001), the present study used TMS to disrupt right vlPFC. Application of TMS to vlPFC, versus to a control site, was expected to increase failure to resist interference from own perspective. Following TMS to vlPFC, an increased number of errors were expected in ToM states that contain perspectives which are maximally different between self and other, such as false belief. No effect following TMS to vlPFC was expected in ToM states which pose similar behavioural difficulty, but do not comprise incongruence between perspectives, such as avoidance desire (see Hartwright et al., 2012 for further discussion). In order to achieve this, we present a 2x2x2 repeated measures design, where the TMS Site (vlPFC/control), and the Belief (true/false) Desire state (approach/avoidance) of an agent was manipulated. A statistically significant TMS Site by Belief interaction, where the effect of TMS to vlPFC induced greater errors and response latencies in false belief reasoning, would support a causal role for vlPFC in the inhibition of self perspective.

METHOD

Participants

Twenty one right-handed adults (9 female; age range 19-28, $\overline{X}$ age = 22 years) completed all sessions of the TMS experiment. All were recruited through the University's research participation scheme. All were given a safety information booklet regarding TMS and MRI prior to participating in the study and gave informed consent in line with the University of Birmingham research ethics. Each was paid a small honorarium for their participation.

Materials and Procedure

Pre-screen

A pre-screen was carried out prior to data collection in order to identify suitable participants. Suitability was determined on the basis of a TMS-MRI safety screening questionnaire (see Appendices 1 and 2) and their ability to perform the Game Show experiment. Participants completed a computer-based, interactive training session that outlined the Game Show experiment then completed two practice blocks. Only participants who could perform above chance, at $p < 0.05$, in the practice blocks of the Game Show experiment were invited to participate in the TMS experiment (see Appendices 1 and 2 for detailed participant screening information). Of twenty seven prospective participants, 6 individuals (4 female; age range 20-39, $\overline{X}$ age = 24 years) were unable to perform the Game Show experiment to above chance and thus did not participate in any of the TMS experiments.

Game Show Experiment

The game show experiment was adapted from a paradigm devised by Apperly et al. (2011) and Hartwright et al. (2012). The task comprised an orthogonal design where the Belief (true (B+) or false (B-)) and Desire state (approach (D+) or avoid (D-)) of a protagonist was systematically manipulated. This resulted in four equally occurring conditions B+D+, B+D-, B-D+, B-D-. Immediately prior to applying the TMS, participants completed 8 practice trials. None of these practice trials were used in the TMS experiment.

The task required participants to watch a sequence of events on a computer and identify whether the protagonist, Simon, would feel happy or sad about the outcome of a game show (Fig. 21). For each round of the game show, Simon was told what prize was on

offer and shown two closed boxes. One box was always empty and the other always contained the prize, but Simon was blind to the contents of both. One box was opened at random and he would win the contents. Therefore, if the box was empty, he would win nothing for that round of the game show. Importantly, as Simon could only win a set number of prizes, he would not always want to win the prize as they varied in desirability, so an empty box could result in Simon being happy.

A single trial consisted of 5 events. The first event, a prize statement, was always displayed at the beginning of the sequence. Next followed three statements, shown singularly and in a random order. These were a belief, a desire and a reality statement (see Fig. 21B). The final event was always a response probe. This depicted Simon looking at the box that had been selected, which was either open or closed. Each experimental block contained 16 trials of interest. In these trials, the box was closed, meaning that Simon could not yet see the contents of the box. Here, participants gave a left/right button response indicating whether they thought Simon was happy or sad about the box that had been selected, on the basis of his belief-desire state regarding what it contained. A further 16 trials, herein termed 'fillers', were devised to prevent formulaic response preparation (see Hartwright et al. (2012) for further discussion). In filler trials, the box was open in the response probe, meaning that Simon could see the contents. Consequently, in this instance, participants needed to indicate whether Simon was happy or sad about the outcome of the game, on the basis of him knowing what the box contained and his desire to win the contents.

The participants completed 4 blocks of the game show task. Each block contained 16 trials and 16 fillers, with an equal number of instances of each belief-desire state. This resulted in 64 trials of interest per participant, comprising 16 of each belief-desire condition, plus 64 fillers. Each trial lasted 12500 ms and was followed with a variable rest period (range

= 2000-3000 ms, mean = 2500 ms) which consisted of a blank screen. Presentation software (v. 14.1; Neurobehavioral Systems, CA) was used to randomise the presentation of trials, present the stimuli and record the behavioural response data simultaneously. The left/right presentation of the response text ‘happy’ ‘sad’ was consistent.

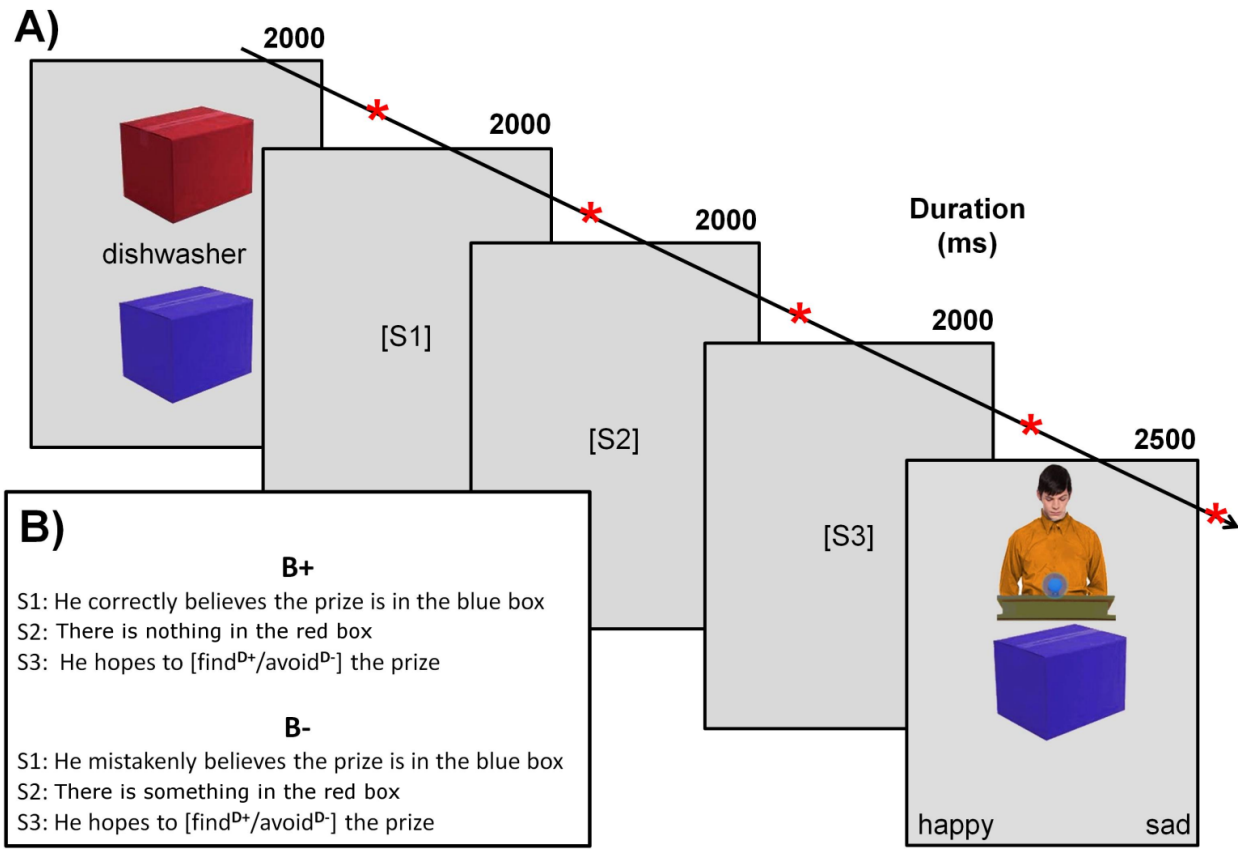

Fig. 21. Method: Game Show Task Paradigm

(Panel A) Schematic example of a single trial. The vertical presentation of the box colour (red/blue) was randomised. The red star * indicates a blank screen shown for 500ms to reduce eyestrain. The colour (red/blue) of the final box in the sequence was randomised. In filler trials, the final element in the sequence depicted an open, rather than a closed, box. No other differences existed (see final event in Fig. 22. for an example). **(Panel B)** Example statements for true (B+) and false (B-) belief scenarios. The temporal order of these statements was randomised. Where text is written within [ ], this denotes that the statement would contain only one of those options, dependent on whether the trial was an approach (D+) or avoidance desire (D-) condition; e.g., He hopes to avoid the prize.

Control Task

To demonstrate that any effect of TMS was specific to the process of interest, a control task was devised using the stimuli from the main experiment (Jahanshahi & Rothwell, 2000). In this computer based task, participants used the buttons on a computer mouse to state whether the image on screen showed an open or closed box (Fig. 22). Thus, the task comprised a repeated measures design where the Box State (open/closed) was systematically manipulated. A single trial comprised an upper-centre justified image depicting Simon looking at the game show podium. An image of either a red or blue box, which was open or closed, was shown mid-centre of the display. These images were exactly as the response probes used in the game show task. The lower left and right quadrants of the screen displayed the words 'open' or 'closed' respectively, to prompt a left/right button response. The left/right presentation of the response text was consistent. Participants were encouraged to respond as quickly and accurately as possible upon seeing the image of the box. Each trial lasted for 2500 ms and was followed with a variable rest period (range = 2000-3000 ms, mean = 2500 ms), which comprised a blank screen. These timings replicated those from the game show task. Participants completed 4 blocks of the control task, which contained 8 repetitions of each colour (red/blue) and box state (open/closed) combination, resulting in 32 trials per block and 128 trials in total. The presentation of each trial type was randomised in real time by the display software, Presentation (v. 14.1; Neurobehavioral Systems, CA).

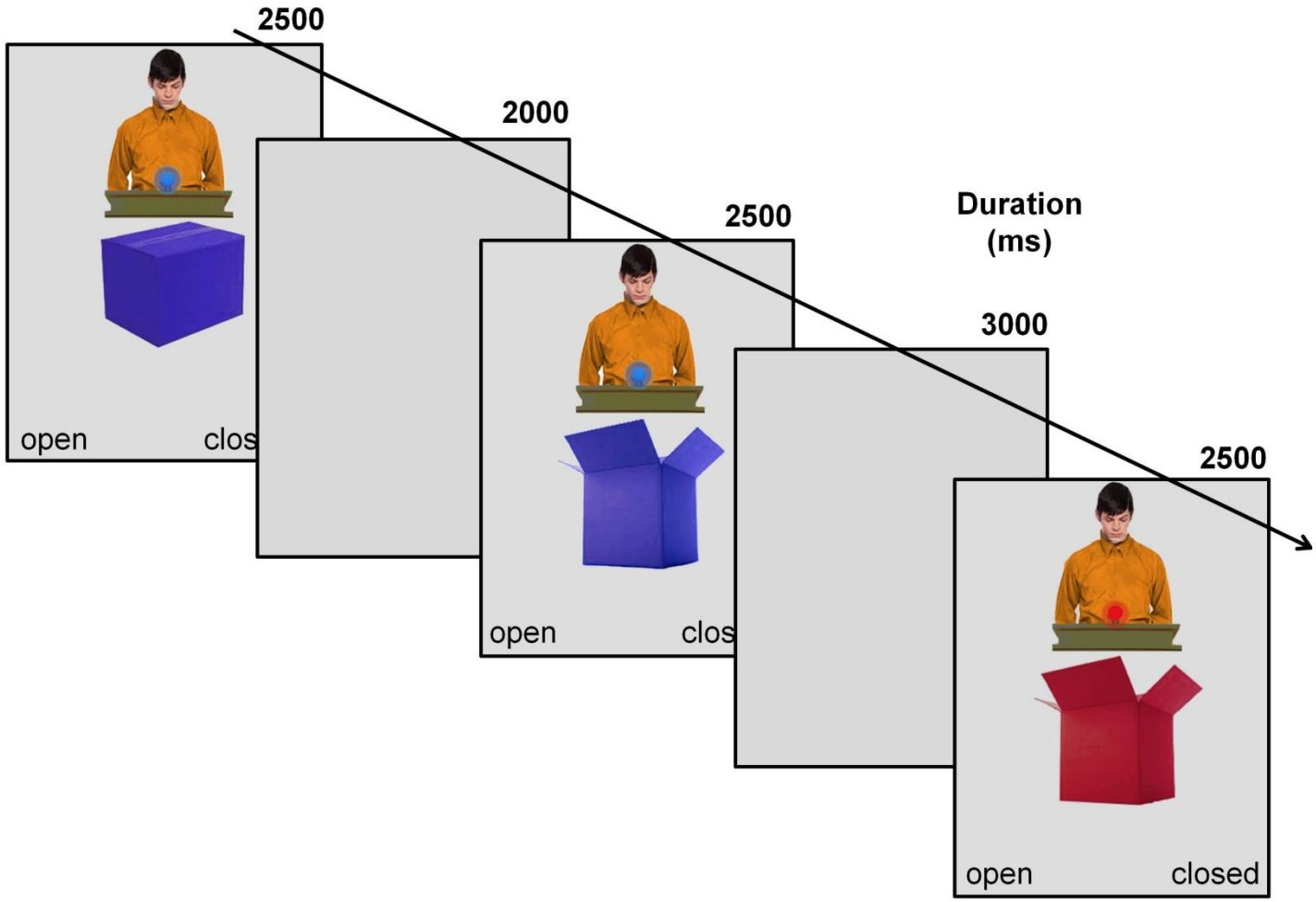

Fig. 22. Method: Control Task Paradigm
Schematic example of three trials. Each trial is followed by a variable rest period, in which a blank screen is shown. The colour (red/blue) and state of the box (open/closed) was randomised.

TMS Procedure

Prior to completing the first TMS session, a T1-weighted anatomical magnetic resonance (MR) image was acquired from each participant using a 3T Philips Achieva scanner. The participants then completed two TMS sessions, conducted over separate days. Each participant was assigned at random to receive TMS to either the site of interest – right vlPFC – or a control site around the vertex – Cz – first. All participants completed both target and control site sessions. The target site was described to participants as a 'frontal site' whereas the control site was described as a 'parietal' site. Participants were blind to the site of interest. Right vlPFC was identified using a group derived coordinate set from a previous ToM experiment conducted with different participants (Hartwright et al., 2012). The Montreal Neurological Institute (MNI) coordinates, [50, 20, -6], were transformed into individual coordinate sets in the current participant group using a series of transformation matrices. The target site was then marked on each participant's anatomical image using Brainsight 2, a system for frameless stereotaxy (v2.2; Rogue Research inc, Canada). The control site, Cz, was identified using skull landmarks and labelled with skin markers. A cTBS paradigm – as outlined in Huang et al. (2005) and Verbruggen et al. (2010) – was administered using a Magstim Rapid2 system (The Magstim Company, Whitland, UK) with a 70 mm figure-of-eight coil. The output intensity was calibrated according to a proportion of the resting motor threshold, range 30-35% of maximum stimulator output. During stimulation, participants were seated with their chin lowered onto a padded rest. A padded block was placed at the left side of the head to minimise movement. For both TMS sites, participants completed 4 blocks of the game show task, which were interleaved with 4 blocks of the control task. The order of task presentation was counterbalanced, so that half of the participants completed a block of

the control task first, followed by the game show task. Behavioural testing following application of the TMS was completed in less than 45 minutes.

RESULTS

Game Show Experiment

The error data were input into a repeated measures ANOVA, with Belief (B+/B-), Desire (D+/D-) and TMS Site (vlPFC/control) as within subjects factors. There was no effect of Belief ($F(1,20) = 2.17$, $p = 0.16$, $\eta^2 = 0.10$) or TMS Site ($F(1,20) = 1.77$, $p = 0.20$, $\eta 2 = 0.08$). A significant main effect of Desire was identified, where error rate in D- > D+ ($F(1,20) = 9.00$, $p < 0.01$, $\eta^2 = 0.31$). No significant interactions between any of the factors were identified, including TMS Site by Belief ($F(1,20) = 3.454$, $p = 0.078$, $\eta^2 = 0.147$). Fig. 23A illustrates the mean proportion of incorrect responses, across each condition and site.

Reaction times (RTs) were recorded from the onset of the response probe until a response was made. Only correct responses were used for RT analysis. Any data points that were 3 standard deviations outside of the participant's condition mean, per TMS site, were considered anomalous and removed. This resulted in the exclusion of 0.1% of responses. A 2x2x2 repeated measures ANOVA was conducted on the remaining data, with Belief (B+/B-), Desire (D+/D-) and Site (vlPFC/control) as within subjects factors. This revealed significant main effects of Belief, where B- > B+ ($F(1,20) = 41.50$, $p < 0.001$, $\eta 2 = 0.68$) and Desire, where D- > D+ ($F(1,20) = 86.74$, $p < 0.001$, $\eta 2 = 0.81$), but no effect of TMS Site ($F(1,20) = 0.50$, $p = 0.49$, $\eta 2 = 0.02$). No statistically significant interactions were identified, including TMS Site by Belief ($F(1,20) = 0.06$, $p = 0.811$, $\eta^2 = 0.003$). Fig. 23B summarises the mean RT for correct responses given across the four conditions, grouped by TMS Site.

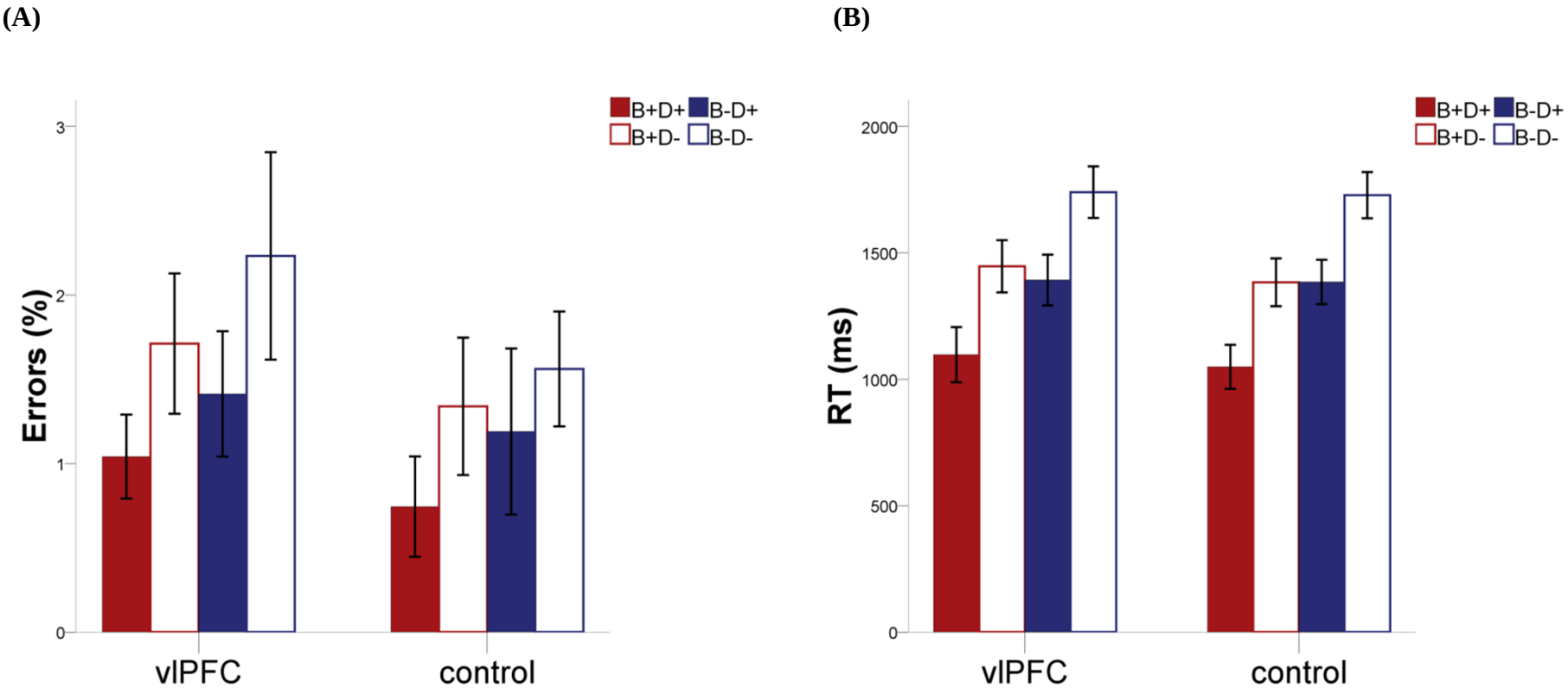

Fig. 23. Results: Game Show Task Behavioural Data
Error bars reflect +/-1 SE of the mean. **(Panel A)** Mean percentage of errors made within condition and site **(Panel B)** Mean RT for each condition and site in milliseconds.

Control Task

The error data were input into a 2x2 repeated measures ANOVA, with site (vlPFC/control) and box state (open/closed) as within subjects factors. There was no effect of TMS Site ($F(1,20) = 0.04$, $p = 0.84$, $\eta 2 = 0.002$); however, a significant main effect of Box State existed, where more errors were made identifying a closed, rather than an open, box ($F(1,20) = 7.33$, $p = 0.01$, $\eta 2 = 0.27$). An interaction between Box State and TMS Site identified that the difference in error rate between open and closed box trials was far greater following TMS to vlPFC than after TMS applied to the control site ($F(1,20) = 6.81$, $p < 0.05$, $\eta 2 = 0.25$). Fig. 24A illustrates the mean proportion of incorrect responses, within each condition and site.

RTs were treated as per the game show task, with outliers being removed prior to RT analysis. This resulted in the exclusion of 1.82% of responses. A repeated measures ANOVA with TMS Site (vlPFC/control) and Box State (open/closed) as within subjects measures found no effect of TMS Site on participant RT ($F(1,20) = 0.004$, $p = 0.95$, $\eta 2 = 0.000$), but a significant main effect of Box State, where closed boxes took significantly longer to identify than open boxes ($F(1,20) = 39.96$, $p < 0.001$, $\eta 2 = 0.67$). No interaction between TMS Site and Box State was identified in RT. Fig. 24B summarises the mean RT for correct responses in open and closed box trials, grouped by TMS Site.

(A) **(B)**

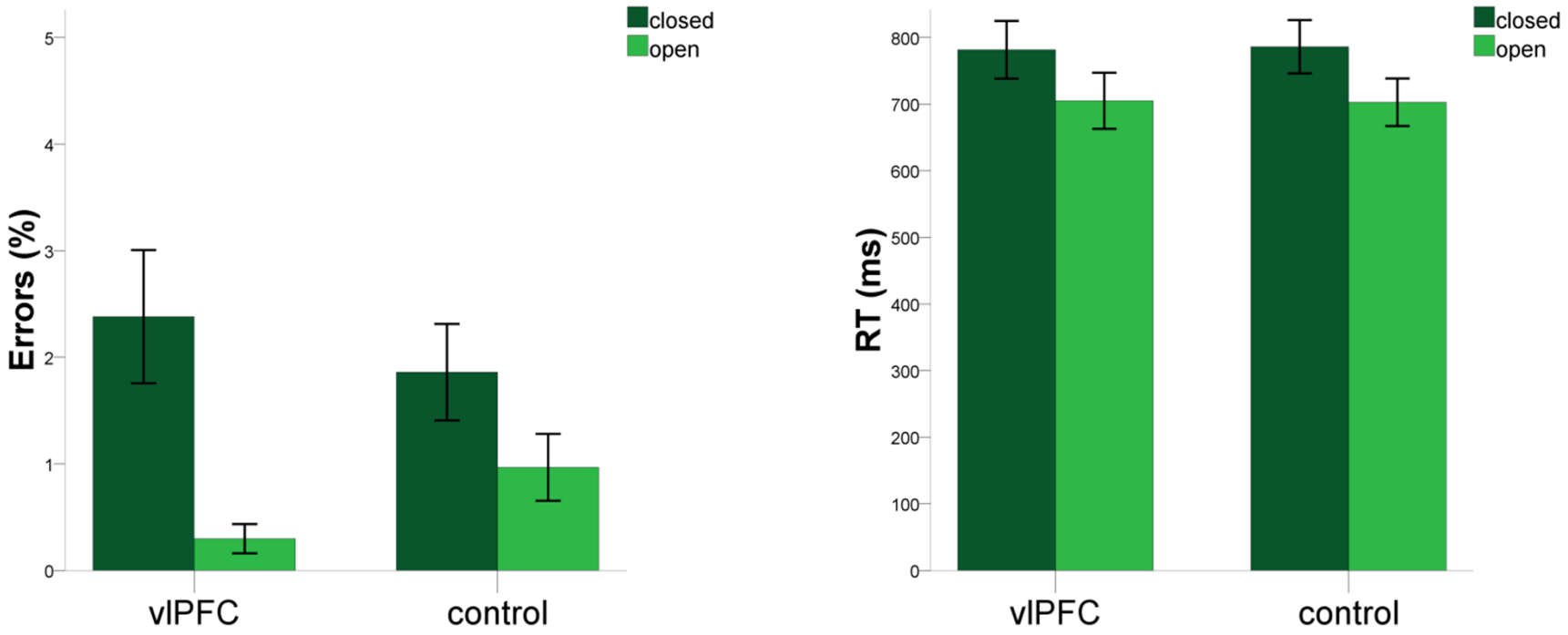

Fig. 24. Results: Control Task Behavioural Data
Error bars reflect +/-1 SE of the mean. **(Panel A)** Mean percentage of errors made within condition and site **(Panel B)** Mean RT for each condition and site in milliseconds.

DISCUSSION

The present study sought to identify a causal role for vlPFC in inhibition of self perspective. By applying cTBS, a TMS protocol thought to depress cortical excitability (Huang et al., 2005; Ridding & Ziemann, 2010), we anticipated that stimulation to vlPFC would affect those ToM states which feature incongruence between self and other, such as false belief, but that it would not impact on other behaviourally difficult ToM states, like avoidance desire. In a three-way ANOVA comparing the factors of TMS Site (vlPFC/control), Belief (true/false) and Desire (approach/avoid), a statistically significant interaction between site and belief would support this hypothesis, where a larger behavioural cost was expected to be associated with TMS to vlPFC in false belief reasoning specifically. Consistent with previous work examining belief and desire reasoning (Apperly et al., 2011; German & Hehman, 2006; Hartwright et al., 2012), the current study identified a behavioural cost for reasoning about an agent who held a false belief or an avoidance desire; however, no overall effect of TMS on ToM was identified in either error rate or RT. Moreover, no interaction between site and belief were identified, indicating no specific effect of TMS to false belief reasoning. The present experiment, therefore, failed to identify a causal role for vlPFC in ToM. Still, on the basis of the data reviewed earlier, further examination is warranted, particularly as neuroimaging data (e.g., Hartwright et al., 2012; van der Meer et al., 2011; Vogeley et al., 2001) converge with lesion data in pinpointing vlPFC in inhibition of self perspective (Samson et al., 2005). The absence of an effect of TMS suggests that this may reflect issues with localization or 'neural efficacy' (see de Graaf & Sack, 2011 for a general discussion), each of which are reviewed in turn.

Localization of vlPFC was guided by individual structural imaging data, where TMS was applied to a group-derived coordinate set from a different sample (Hartwright et al.,

2012). Unlike phosphenes, which can be induced for early visual areas, and muscle twitches for primary motor cortex, cognitive processes involve areas of silent association (de Graaf & Sack, 2011; Jahanshahi & Rothwell, 2000). It is therefore not possible to be certain that the coordinate set chosen represented the foci of cortex responsible for self perspective inhibition in the sample here. However, approximate convergence with other published data supports the view that the coordinates selected are representative of vlPFC activation in ToM (e.g., see Spreng et al., 2009). Nonetheless, although TMS was applied using frameless stereotaxy to ensure accurate localization, due to large muscle masses around the prefrontal cortex, subject movement did occur during stimulation. Despite padding to stabilise the head, the vast majority of participants displayed activation of the trigeminal nerve, resulting in jaw clenching and mouth twitching, and the supraorbital nerve, causing a blink reflex, at each pulse. Note that this is not unusual during stimulation to this region (Jahanshahi & Rothwell, 2000). Maximum coil displacement during stimulation to vlPFC ranged from 0.8-3.0 mm (mean = 1.48 mm, *SD* = 0.55); thus, it could be argued that stimulation may have been applied to neighbouring cortex, which may not support the process of interest. Still, as the coil placement was adjusted in real-time during stimulation, any deviation from the target site would have been momentary.

An alternative possibility for the null result relates to 'neural efficacy' (de Graaf & Sack, 2011). That is, that the TMS protocol used was not sufficient to disturb processing in vlPFC, or, that the protocol did, in fact, sufficiently disrupt processing, but that neighbouring regions were able to compensate. The former argument is unlikely, given that vlPFC impaired Box State judgement in the control task. The latter is difficult to determine without having taken some additional measurement at the time, such as electro- or magneto-encephalography (EEG/MEG respectively). Either way, Hartwigsen et al. (2013) demonstrated that cTBS to left

inferior frontal gyrus (IFG) produces a facilitatory drive from the homologous region to the lesioned area, resulting in shorter response latencies. Thus, short term cortical reorganisation caused by cTBS shows that the contralateral region can be drafted to support the primary functions of the initial cite of perturbation. Nevertheless, there is, as yet, no conclusive evidence that this effect may apply to the right vlPFC.

Lastly, our results suggest that consideration should be given to recent work which calls into question the overall efficacy of the cTBS protocol (Hamada, Murase, Hasan, Balaratnam, & Rothwell, 2012; McAllister et al., 2013; Vernet et al., 2013). For example, using MEG, McAllister et al. (2013) examined the neuroplastic effects of cTBS on M1. Notably, only half of the participants tested, herein termed 'responders', showed a significant decrease in corticospinal excitability. The remaining participants, classed as 'non-responders', either showed no effect or a marginal increase. In addition, the effect on simple motor response latencies differed according to corticospinal excitability; whilst responders were significantly slower to respond following cTBS, non-responders showed a decrease in RT. Similarly, Hamada et al. (2012) were only able to identify a quarter of individuals who responded as expected to TBS. Although this variability refers to cTBS applied to the motor cortex in both studies, it is plausible that a similar issue would apply to other cortical areas. The paucity of published null results, however, makes it difficult to determine if this is the case (although see Verschuere, Schuhmann, & Sack, 2012). Either way, the factors which influence the induction of plasticity are not fully understood; susceptibility is likely influenced by factors such as age, attention, sex, genetics and time of day (Ridding & Ziemann, 2010), adding further noise to the data. It should be reiterated that the control task data suggest that the current sample were susceptible to some sort of effect of cTBS to vlPFC, although post-hoc evaluation of the behavioural data suggest a distinct divide within the

sample, where over 40% of the participants actually became faster at responding following TMS to vlPFC. Of those, over half were less error prone, or showed no change in error rate following TMS to vlPFC, suggesting this subgroup were not simply pressing the button faster. Of course it may be that, like Hartwigsen et al. (2013) demonstrated in left IFG, temporary cortical reorganisation from cTBS caused the contralateral homolog to support the functions served by our region of interest in this subgroup of participants; thus, cancelling out any opposite effect in the non-responders. The sample size of possible responders and non-responders in the current study, however, does not permit any statistically informed conclusion here. Regardless, like Hamada et al. (2012); McAllister et al. (2013), our data suggest that the effect of cTBS is highly variable across subjects, making it difficult to draw any meaningful conclusion regarding the role of vlPFC in ToM with the current data.

In sum, the present study used a cTBS protocol to identify a causal role for vlPFC in ToM; specifically, in inhibiting self perspective in those mental states where the knowledge state is incongruent between self and other. The behavioural data were consistent with the existing literature, in that a cost was associated with holding in mind a false belief or an avoidance desire. However, no effect of TMS was found on ToM. The hypothesised interaction between TMS Site and Belief state, where an increase in errors and response latencies was expected in false belief following TMS to vlPFC, was not identified. Lesion data from Samson et al. (2005) provides compelling evidence that vlPFC is recruited to inhibit a prepotent self perspective: impairment in the ability to resist interference from self perspective was demonstrated to coincide with a lesion to the right frontal cortex. When considered with a powerful interaction effect identified in a proximal region using fMRI (Vogeley et al., 2001), and further converging neuroimaging data (Hartwright et al., 2012; van der Meer et al., 2011), it is difficult to dismiss a role for vlPFC in ToM on the basis of the

present result, particularly as recent studies now suggest that cTBS does not produce consistent effects across subjects (Hamada et al., 2012; Hartwigsen et al., 2013; McAllister et al., 2013; Vernet et al., 2013).

CHAPTER 5:

THE SPECIAL CASE OF SELF PERSPECTIVE INHIBITION IN MENTAL, BUT NOT NON-MENTAL, REPRESENTATION[7]

[7] This chapter is currently in preparation for submission: Hartwright, C. E., Apperly, I. A., and Hansen, P.C. (in prep.). *The special case of self perspective inhibition in mental, but not non-mental, representation.* Manuscript in preparation.

ABSTRACT

The ventrolateral prefrontal cortex (vlPFC) has been implicated in studies of both executive and social function. Recent meta-analyses suggest that vlPFC plays an important but little understood role in Theory of Mind (ToM). Converging neuropsychological, functional Magnetic Resonance Imaging (fMRI) and event related potential (ERP) data, suggest that this may reflect inhibition of self perspective. The present study adapted an extensively published ToM localizer to evaluate the role of vlPFC in inhibition of self perspective. The classic false belief, false photograph vignettes that comprise the localizer were modified to reflect high and low salience of self perspective. Using a factorial design, the present study identified a behavioural and neural cost associated with having a highly salient self perspective that was incongruent with the representational content. Importantly, vlPFC only differentiated between high versus low salience of self perspective when representing mental state content. No difference was identified for non-mental representation. This result is explored in terms of the different processes that may be required to represent competing mental, and non-mental content.

INTRODUCTION

The frontal lobes have long been viewed as critical for supporting complex social cognition. From comparative and neurodevelopmental studies, through to lesion and neuroimaging data, the co-localization of social and executive processes to frontal cortex illustrates this region's role in supporting complex higher functions (e.g., Amodio & Frith, 2006; Duncan & Owen, 2000; Eslinger, Grattan, Damasio, & Damasio, 1992; Ridderinkhof, Ullsperger, Crone, & Nieuwenhuiss, 2004; Rowe, Bullock, Polkey, & Morris, 2001;

Semendeferi, Lu, Schenker, & Damasio, 2002; Stone, Baron-Cohen, & Knight, 1998; Stuss & Benson, 1984; Stuss, Gallup, & Alexander, 2001). As cognitive neuroscience has begun to home in on the neural correlates of more specific constituents of social cognition, Theory of Mind (ToM), the ability to attribute mental states, thoughts and intentions to others, raises some interesting challenges in terms of understanding the interplay between the neurocognitive bases of social processes and any supporting executive functions. One such challenge concerns successfully identifying neural regions which support the expression of ToM, for example, in terms of executive processes.

Considerable effort has been directed towards attributing specific functional profiles to the temporoparietal junction (TPJ) and medial prefrontal cortex (mPFC), regions thought to be critical in ToM (for reviews see Carrington & Bailey, 2009; Lieberman, 2007; Mar, 2011; Spreng et al., 2009; Van Overwalle, 2009). Quantitative meta-analyses, however, suggest consistent recruitment of several, less examined, regions across a multiplicity of paradigms, including the amygdala, precuneus and ventrolateral prefrontal cortex (vlPFC) (see Bzdok et al., 2012; Mar, 2011; Spreng et al., 2009). The vlPFC, particularly left inferior frontal gyrus (IFG), has been described as a possible candidate for part of the "core mentalizing network" (Mar, 2011, p.124); however, until fairly recently, the appearance of vlPFC in ToM has been largely unexamined. Nonetheless, there are good reasons for thinking that this region may serve an important process in ToM.

It is well documented that vlPFC is involved in executive control processes (e.g., Aron, Fletcher, Bullmore, Sahakian, & Robbins, 2003; Aron, Robbins, & Poldrack, 2004; Badre & Wagner, 2007; Garavan, Ross, & Stein, 1999; Konishi et al., 1999). Developmental studies provide considerable evidence that both children and adults have difficulty with certain ToM states. This is thought to reflect underlying executive control processes, in terms

of suspending self perspective in favour of someone else's, or selecting from competing perspectives (Birch & Bloom, 2004; Birch & Bloom, 2007; Carlson & Moses, 2001; German & Hehman, 2006). For example, in a typical false belief task, considered the litmus test for ToM, a scenario is outlined wherein an agent places an object in a specific location. Unbeknown to the agent, the object is then moved from the original location to a new location. As a consequence, the agent is described as holding a 'false belief', as their belief reflects a misinformed state of reality. In order to demonstrate having a ToM representation of the agent, participants must identify which location the agent will search first in order to retrieve the object (e.g., see Wimmer & Perner, 1983). This can only be solved by setting aside knowledge of reality and selecting the location where the agent originally left the object. This conceptual analysis highlights that there are two information streams which may jostle for selection: 1) the new, true location of the object and 2) where the agent (incorrectly) believes the object is. Thus, successful expression of the agent's ToM in this unexpected transfer task will require some form of control process, which would explain the behavioural difficulty that people demonstrate.

One approach to identifying neural regions which might support control processes in ToM is to compare the neural correlates of assigning a false, versus a true (i.e. reality congruent) belief to an agent. In doing so, assigning a true belief would be expected to attract lesser inhibitory demands as, in the unexpected transfer task, there is no difference between the believed and the true location. Whilst there are still two information streams to keep in mind – what the agent believes and what the real state of affairs is – as these are congruent, any interference between the two is likely to be low compared with false belief. Of the few studies that have compared true and false belief reasoning, these converge on lateral PFC and anterior cingulate cortex (ACC) for false, over true, belief reasoning (Döhnel et al., 2012;

Hartwright et al., 2012 [Chapter 2 here]; Sommer et al., 2007). This result is further refined by ERP data, which suggest that lateral frontal activity, in particular, reflects inconsistency between perspectives, for example, when judging own visuospatial perspective against that of an agent (McCleery et al., 2011). Whilst comparing true with false belief provides one method of varying conceptual perspective differences between self and other, this approach is not unproblematic. First, it is difficult to be certain that the status of an agent's belief, and not the real state of affairs, is being represented (Aichhorn et al., 2009; Hartwright et al., 2012 [Chapter 2 here]). Second, as well as manipulating inhibitory demands, such an approach also systematically varies the truth value of the belief. An alternative approach, where the salience of one's own perspective is manipulated, however, would not confound truth status with the requirement to inhibit self perspective.

Vogeley et al. (2001) provide an early attempt to manipulate the salience of own conceptual perspective whilst adopting the perspective of someone else. A series of short vignettes described the participant as a central agent in the story, for example, as the owner of a shop that has just been burgled. This approach meant that the participant had to ascribe the behaviour, attitudes and perceptions of this central agent – in this instance, *'you'* the shop owner – to themselves; therefore, raising the salience of their own perspective. An alternative series of vignettes replicated the structure and style of the high salience vignettes, but the participant did not feature as a character in the story. In this case, then, there was just *'the'* shop owner, therefore making the participant's own perspective less salient. As well as systematically manipulating the presence of self in these vignettes, the authors also varied the presence of another agent's behaviour, attitudes and perceptions. The resulting design comprised four conditions where presence of Self (absent/present) and presence of ToM (absent/present) were manipulated. Variation in Self was shown to modulate bilateral ACC,

TPJ and precuneus. Variation in the ToM also included medial frontal activation, spanning dorsal regions and ACC, as well as bilateral lateral PFC. Although not the primary focus of the study, Vogeley et al. (2001) identified that a single, small area of right IFG was recruited for the interaction between ToM and Self, that is, when participants were required to feature as an agent in the story, whilst making a ToM judgement about a further character in the story. The authors suggested that this activation may have reflected an additional executive process, which was required in the instance of taking someone else's perspective, whilst having to integrate this with their own perspective.

The interaction effect identified in Vogeley et al. (2001) was further clarified in a single case study presented by Samson et al. (2005). Following damage to right lateral PFC, encompassing the region of IFG identified by Vogeley et al. (2001), patient WBA demonstrated impairment in high, but not low, inhibition false belief reasoning. The high inhibition false belief task, which is a more typical ToM paradigm, comprised a short non-verbal film depicting the unexpected transfer task outlined earlier. The footage showed a woman watching a man place an object into one of two containers. She then leaves the room. Whilst the woman is away, the man takes the object and places it into the other container. Importantly, the footage clearly shows which of the two containers the man places the object into. The woman then returns to the scene. The patient's task was to point to which one of the two containers he thought the woman would look in first to find the object. In order to successfully identify the correct container, the patient needed to infer that the woman held a false belief, but also inhibit his own knowledge of the true location of the object. On this task, patient WBA failed to select the correct container in all but one of the trials: he was systematically selecting the location that conferred to his own knowledge point.

In a further experiment, Samson et al. (2005) tested patient WBA on a similar paradigm, but with reduced inhibitory demands. As with the first task, each trial of interest comprised a non-verbal film depicting the unexpected transfer of an object. Again, the footage showed a woman watching a man place an object into one of two containers. Crucially, this time, it was not possible to see which container the object was hidden in. The woman then leaves the room and, whilst she is away, the man swaps the two containers. She then returns and points to one of the containers to provide a clue as to where she believes the object is hidden (i.e. she points to its initial, but now incorrect, location). The patient's task was to identify which of the two containers really housed the object at that point in time (i.e. point to its new location). In order to correctly identify where the object was, the patient needed to infer that the woman held a false belief as to the object's location, and consequently point towards the opposite box that she had hinted at. As the patient was never shown the true location of the object, the patient did not need to inhibit his own knowledge, in order to infer the woman's false belief. In a striking contrast to the first experiment, patient WBA selected the correct container in all but one trial: he successfully adopted the view point of an external agent. When his performance in the two tasks are considered together, damage to the right lateral PFC appeared to cause an inability to resist interference from his own perspective, not a ToM deficit as such.

In an attempt to further understand the form of inhibitory processes in conceptual perspective taking, to our knowledge, only two further neuroimaging studies have been conducted. The first used a modified version of the high and low inhibition tasks in Samson et al. (2005). Here, van der Meer et al. (2011) collected fMRI data from neurologically intact adults whilst they watched high versus low inhibition false belief scenarios. The same participants also completed a classic Go/No-Go task. In high versus low inhibition scenarios,

frontal activation was limited to bilateral vlPFC and dorsal mPFC. Similarly, No-Go versus Go trials elicited bilateral vlPFC. Common to high > low inhibition and No-Go > Go was left lateral PFC and right vlPFC. These data led the authors to conclude that inhibition of own perspective is mediated by bilateral vlPFC when supporting a functioning ToM.

Along a similar vein, Rothmayr et al. (2011) asked participants to identify whether an agent, on the basis of their false belief, looked for the transferred object in an expected or an unexpected location. They used the same pictorial stimuli to create a separate, novel Go/No-Go task. Contrast masking analyses identified that a largely left lateralised network, including left IFG and the wider lateral PFC, was recruited exclusively in false > true belief versus No-Go > Go trials. A conjunction between the false > true belief and No-Go > Go identified right dorsal mPFC and dorsolateral PFC bilaterally, plus bilateral TPJ and other regions outside of the PFC. On the basis of common neural recruitment during the ToM and inhibitory control tasks, and other research examining either false belief or inhibitory control, the authors conclude that TPJ, dorsal medial- and lateral PFC support domain general processes common to both ToM and executive control[8]. What is particularly interesting here, however, is that left IFG responded preferentially to conflict in ToM, over a more classical motor-inhibition task. In line with Mar (2011), Spreng et al. (2009), Samson et al. (2005) and Vogeley et al. (2001), this provides a further hint that IFG serves a selective role in ToM.

Whilst a role for vlPFC in ToM emerges from the existing literature, little is known about how this region responds to fine-grained variation in perspectives, particularly within mental and other, structurally matched, non-mental representation tasks. The present study therefore sought to examine the role of vlPFC, specifically in the inhibition of self perspective during ToM and physical representation. The present study comprises a simple manipulation

[8] Although, note they place greatest emphasis on the role of dorsal mPFC

to an extensively published ToM localizer task, created by Saxe and Kanwisher (2003) (e.g., see Aichhorn et al., 2009; Hartwright et al., 2012 [Chapter 2 here]; Mitchell, 2007; Perner et al., 2006; Saxe & Powell, 2006; Saxe & Wexler, 2005; Scholz et al., 2009). This localizer task was used as it controls for some of the attentional differences that would exist if comparing false versus true belief, for example. By modifying this task to include vignettes which feature high and low salience of self perspective, the present study examined the effect of varying inhibitory demands in mental (ToM, i.e., false belief) versus non-mental (physical, e.g., false photograph, false drawing etc) representation. Following on from the quantitative reviews by Spreng et al. (2009) and Mar (2011), alongside neuropsychological evidence from Samson et al. (2005), ERP data from McCleery et al. (2011) and neuroimaging data from Vogeley et al. (2001), Rothmayr et al. (2011) and van der Meer et al. (2011), the analyses focused on vlPFC. This enabled close inspection of the functional profile of this region in response to high, versus low, self perspective inhibition in conceptual ToM and non-ToM representations. On the basis of the prior neuroscience research reviewed, it was anticipated that vlPFC would be modulated on the basis of high versus low salience of self perspective, as a result of inhibitory demands. From the behavioural literature, an effect of this manipulation may be elicited in terms of error rates, where interference from self perspective may increase the number of errors when making a representation that featured a highly salient self perspective (Birch & Bloom, 2004; Birch & Bloom, 2007; Carlson & Moses, 2001; German & Hehman, 2006). The nature of the localizer, however, led us to this make this assertion tentatively as, for typical adults, it is a relatively straightforward, slow-paced task. Consequently, the behavioural measure that can be obtained from this task may not be sensitive enough to identify such subtle effects (Birch & Bloom, 2004). There is little information regarding whether cognitive load from competing perspectives in conceptual

perspective taking will differ between mental versus non-mental representation tasks; however, on the basis that self perspective is present in both false belief and false photograph scenarios (Hartwright et al., 2012, [Chapter 2 here]), and, consistent with this expectation, the original localizer does not recruit lateral PFC, we anticipated that vlPFC will not differentiate between representational tasks.

METHOD

Participants

Twenty one right-handed, neuro-typical adults (12 female; age range 19-28, $\overline{X}$ age = 22 years) participated in exchange for a small honorarium. All were recruited through the University's research participation scheme and gave informed consent in line with the University of Birmingham research ethics. The Wide Range Achievement Test – Third Edition (WRAT-3) Reading Scale was administered prior to taking part in the experiment to ensure reading proficiency commensurate with the task (see Appendix 1 for detailed participant screening information).

Materials and Procedure

Modified Theory of Mind Localizer Experiment

The task was based on a localizer procedure devised by Saxe and Kanwisher (2003). Stimuli were based on a modified and expanded selection of the localizer stories (Saxe & Andrews-Hanna, n.d.). All of the vignettes were rated for ease of understanding and trialled on a separate group of individuals prior to running the fMRI experiment. For the fMRI

experiment, participants read a total of 56 short vignettes, which referred to either a mental representation (belief), such as an agent with a false belief, or a non-mental representation (physical), such as a false photograph, video or painting (see Zaitchik (1990)). For the sake of simplicity, and following previous terminology for stimuli of this nature, the physical representation stimuli are herein referred to as 'false photographs'. The important feature of these vignettes is that they refer to a change of state which causes incongruence between reality and the representational content. Thus, in both types of vignettes, the term 'false' is used to illustrate that the representational content of is outdated and, therefore, no longer in line with reality.

In order to modulate neural regions that support self perspective inhibition, following a theoretically similar approach to Samson et al. (2005), the original vignettes were modified so that the salience of the participant's perspective – the information outlining the true state of affairs – was systematically varied from high to low. Each belief vignette and each physical vignette had a high and low salience version. In high salience vignettes, the occurrence and precise nature of the change of state was made explicit; thus, the viewer held a highly salient perspective of reality. The resulting striking incongruence between own perspective and the representational perspective was anticipated to cause considerable interference when adopting the false, representational perspective, resulting in behavioural and neural consequences (Birch & Bloom, 2004; Birch & Bloom, 2007). In low salience vignettes, the occurrence of the change of state is still made explicit; however, the precise nature of reality is under specified, making the viewer's perspective of the true state of affairs less salient. It is important to highlight that incongruence between the representational and real state of affairs is again present although, this time, incompatibility between the participant's perspective and the state of reality is less vivid. As a consequence, it is expected that participants will suffer

less interference from their own knowledge point when adopting the false, representational perspective.

The first example illustrates the format of a mental representation vignette with high self perspective inhibition demands,

Lorraine dashed out the door and mistakenly left her lunch money on the side table. Thinking it was for a school trip, her daughter put the money upstairs in her purse.

-

Lorraine expects to find her lunch money in her daughter's purse.

True *False*

In this example, it is clear that Lorraine believes her lunch money is on the side table. From reading this scenario it becomes apparent that, unbeknown to Lorraine, the money was moved to her daughter's purse. In order to successfully identify where Lorraine thinks her money is, the viewer must inhibit the reference to the current situation – *the money is upstairs in her daughter's purse* – whilst framing an answer in terms of the agent's false, i.e. outdated, belief – *Lorraine believes her money is on the side table* (conceptual analysis adapted from Russell, Saltmarsh, & Hill, 1999). In this example, then, the answer is false.

The same format is followed for high salience non-mental representation vignettes. For example,

The speed camera captured an image of the bright red car as it sped along the road. The following day, the car was painted a grey colour and the number plates were changed.

-

According to the speed camera image, the car is bright red.

True *False*

This example makes clear that the speed camera footage captures an image of the bright red car. After reading the scenario it becomes apparent that, after the footage was taken, the car was painted a grey colour. In order to successfully identify which colour the footage shows, the viewer must inhibit the reference to the current situation – *the car has since been painted grey* – whilst framing an answer in terms of the false, i.e. outdated, photographic representation – *the footage shows the car when it was red* (conceptual analysis adapted from Russell et al., 1999). In this example, then, the answer is true.

In low salience vignettes, the occurrence of the change of state is made explicit; however, the precise nature of the true state of affairs is loosely specified. Incongruence between the representational and real state of affairs is again present although, this time, the viewer's perspective of the true state of affairs is less salient. For example, for mental representation,

Liz hurried out the door and mistakenly left her coffee money on the desk. Thinking it was for his school lunches, her son put the money away in the usual safe place.

-

Liz expects to find her coffee money on the desk.

True *False*

In this example, the correct answer is true: it is clear that the Liz will believe that her coffee money is on the desk; however, whilst the viewer knows that the money is no longer on the desk, they are not privy to the precise location of the money. As in high salience vignettes, the viewer follows the same process of inhibiting the real state of affairs and selecting the representational content; however, the viewer is expected to suffer less interference from their own knowledge of reality versus the representational perspective.

Similarly, the final example illustrates the format for a low salience, non-mental representation vignette,

The traffic camera snapped an image of the dark blue car as it jumped the traffic lights. The following day, the car was painted a bright colour and the number plates were changed.

-

According to the traffic camera image, the car is a bright colour.

True *False*

In this last example, the correct answer would be false. Overall, whilst the process of self perspective inhibition is required in both high and low salience vignettes, the interference suffered from competing sources of information is expected to vary as a function of salience.

The paradigm comprised a 2x2 repeated measures design with two within-subjects factors, representation (belief/photo (B/P)) and salience of self perspective (high/low (H/L)), collapsed into four equally occurring conditions: BH, BL, PH, PL. Each vignette was displayed in black Arial point 22 font, presented on a grey background. A single trial comprised a short story, displayed for 10 s, followed for 4 s by a true or false question about the preceding story. This required participants to make a response using a two button box that was placed in their left hand[9], where the left button was always used to indicate a true statement. The experiment contained an equal number of true/false responses, which were randomised across the experiment. Stories alternated between belief and physical and were interleaved with a 13.5 s rest period comprising a fixation dot. The presentation of high versus low salience was pseudo randomised to prevent more than three repetitions of either saliency variant, and pairs of high/low stories were not repeated within the same block. The word lengths of each type of vignette were equivalent (belief versus physical representation $t(23) = 0.073$, $p = 0.943$; high versus low salience $t(23) = 0.000$, $p = 1.000$).

The experiment comprised four blocks of 12 trials of interest, each containing three vignettes of each condition. This resulted in 12 trials for each of the four conditions. Participants completed four practice trials immediately prior to scanning to orientate themselves with the task. Each block also contained a further two randomly placed anti-strategy trials. Whilst the structure of the story element in these vignettes was identical to the trials of interest, the question phase required participants to answer a true/false question about

[9] The left hand was used to differentiate potential language related neural activity from motor activity.

the true outcome described within the vignette. This prevented participants from adopting a formulaic approach to response preparation (Saxe & Kanwisher, 2003). The example below illustrates one such vignette where the answer, on this occasion, would be false.

Expecting the game to be postponed because of the rain, the Jones family left the match early. The score was tied, 0-0. During their journey the rain stopped and the game ended with a score of 5-1

-

The final score was tied at 0-0.

True *False*

Data Acquisition and Analysis

Neuroimaging Data Acquisition and Preprocessing

The data were acquired during a single session using a 3T Philips Achieva scanner, , with an 8 channel head coil. The stimuli were presented using Presentation software (v. 14.1; Neurobehavioral Systems, CA), which also recorded the behavioural response data simultaneously. 159 T2*-weighted echo-planar imaging (EPI) volumes were obtained per block of the experiment, each of which consisting of 42 axial slices obtained consecutively in a bottom up sequence, reconstructed voxel size = $3x3x3mm^3$. Whole brain coverage was achieved with a TR = 2.5 s, TE = 35 ms, acquisition matrix = 96 x 96, flip angle = 83°, SENSE factor = 2, voxel size = $3x3x3mm^3$. High resolution T1-weighted structural images

were acquired following collection of the functional data (3D TFE, sagittal orientation, TR=8.4 ms, TE=3.8, matrix size 288x288, 175 slices, reconstructed voxel size = 1×1×1mm^3).

Preprocessing and statistical analyses of the data were performed using the FMRIB software library (FSL version v.4.1.9; FMRIB, Oxford, www.fmrib.ox.ac.uk/fsl; FEAT version 5.98). Initial preprocessing of the functional data consisted of slice timing and motion correction using rigid body transformations (MCFLIRT). The blood oxygen level dependent (BOLD) signals were high-pass filtered using a Gaussian weighted filter of 21 s. The BOLD data were then spatially smoothed using a 5mm full-width-half-maximum kernel. The functional data were registered to their respective structural images and transformed to a standard template based on the Montreal Neurological Institute (MNI) reference brain, using a 7-DoF linear transformation (FLIRT).

fMRI Data Analysis

Four explanatory variables (EVs) of interest – BH, BL, PH, PL – were modelled to reflect the four experimental conditions. Each EV comprised the story and question phase of a single vignette (14 s). Each EV was convolved with a gamma derived hemodynamic response function (HRF) within a general linear model (GLM) framework. The anti-strategy trials and motion parameters were modelled as regressors of no interest. Higher level modelling was used to aggregate the data across participants within a mixed effects model. For confirmatory purposes, a whole brain analysis was computed to replicate the original localizer contrast belief > physical, as per Saxe and Kanwisher (2003). To address the area of interest for the present study, a series of novel higher level analyses were computed. Given our strong a priori hypothesis, and due to the published statistical challenges in identifying neural regions that

support complex cognitive functions (Lieberman & Cunningham, 2009), pre-threshold masking was used to constrain these analyses to bilateral vlPFC[10]. This was defined as comprising ventrolateral voxels in Brodmann Areas 44, 45 and 47 which had a ≥ 20% probability of falling within either the inferior frontal gyrus pars, the frontal operculum or frontal orbital cortices, as classified by the Harvard Oxford Cortical Atlas. The resulting Z statistic images were then thresholded using a cluster based approach, where $Z > 2.3$, cluster $p_{corr} < 0.05$. The final result reflected a 2x2 repeated measures ANOVA with representation (B/P) and salience of self perspective (H/L) as within subjects factors. To provide directional information regarding the effect of salience on representation, a quadrupled t-test was also computed using the same parameters. Results are described according to gross anatomical regions and anterior, mid and posterior vlPFC, approximating BA 47, 45 and 44 respectively (Badre & Wagner, 2007). Mean percent signal change plots of each effect identified by the GLM analyses were created using FSL's Featquery. This enabled a closer examination of how the data contribute to each result (Poldrack & Mumford, 2009).

[10] All neuroimaging analyses were verified using whole brain analysis. This indicated the likelihood of the presence of type II errors in frontal regions, where smaller clusters were eradicated by cluster thresholding at $Z < 2.3$, $p < 0.05$ in a whole brain approach. By focusing the analyses on a smaller number of voxels, the multiple comparison correction would be less severe. This approach enabled us to apply standard cluster detection, whilst addressing the need to minimise type I and type II errors, in an analysis that was sensitive to activation in vlPFC. Importantly, all results presented were consistent with the raw data at group level. See Souza, Donohue, and Bunge (2009) for a similar approach in vlPFC.

RESULTS

Behavioural Data

The participants' error rates were analysed in a 2x2 repeated measures ANOVA, with representation (B/P) and salience of self perspective (H/L) as within subjects factors[11] (Fig. 25A.). Overall, few incorrect responses were made. Whilst, across the whole experiment, participants made fewer errors when making a belief inference versus when making a physical representation judgment, factorial analysis found no significant effect of representation ($F(1,17) = 0.46$, $p = 0.51$, $\eta^2 = 0.03$). Almost three times as many errors were made when self perspective was highly salient, which was supported by a significant main effect of salience ($F(1,17) = 10.07$, $p = 0.006$, $\eta^2 = 0.37$). Post hoc comparisons confirmed a significant difference in the number of errors made in high versus low salience (high > low; $SE = 0.44$, $p = 0.006$). As illustrated in Fig. 25., participants were generally most error prone when making a physical representation judgment and their self perspective highly salient; however, there was no statistically significant interaction between representation and salience in error rate ($F(1,17) = 2.69$, $p = 0.12$, $\eta^2 = 0.14$). Inspection of the error rate for anti-strategy trials determined that the participants were generally performing at ceiling, and thus attending to the real state of affairs (mean frequency correct 7.23/8. *SD* 0.89).

[11] Due to an equipment failure, behavioural data from 3 participants were unavailable. These participants were included in subsequent neuroimaging analyses as their inclusion yielded no key differences in the localization of effects in the raw, unthresholded data.

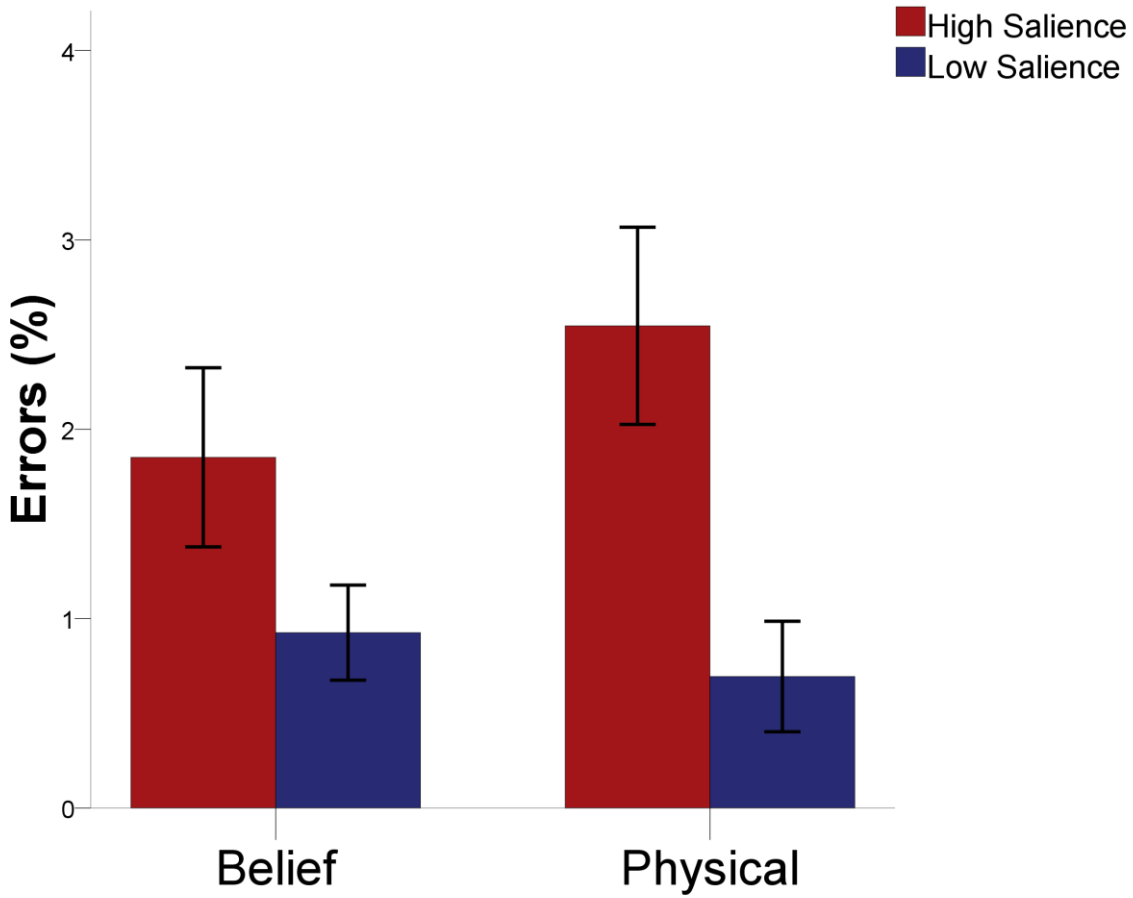

Fig. 25. Results: Modified ToM Localizer Behavioural Data
Mean percentage of errors made per condition. Error bars reflect +/-1 SE of the mean.

Neuroimaging Data

Whole brain analyses of the false belief > false photo contrast, as computed by Saxe and Kanwisher (2003), indentified regions consistent with ToM and previously published uses of the localizer. This included bilateral temporoparietal junction, temporal poles, precuneus and medial prefrontal cortex (Table 6; Fig. 26).

For the main analysis within bilateral vlPFC, a 2x2 repeated measures ANOVA with representation (Belief/Physical) and salience of self perspective (High/Low) as within subjects factors was conducted. This identified a main effect of representation (B/P) in bilateral vlPFC, where the peak activation was centred on mid vlPFC, the inferior frontal gyrus pars triangularis (IFG_{tr}). This cluster extends in the rostral direction to include a small portion of the superior quadrant of anterior vlPFC, and in the caudal direction towards the posterior boundary of the frontal poles. This main effect is driven by a larger percentage signal change in physical versus belief stimuli (Fig. 27A; Fig. 27D. red shading). No main effect of salience of self perspective (H/L) was identified; however, a two-way interaction identified a separate cluster which chiefly comprised left anterior vlPFC (Fig. 27B; Fig. 27D. blue shading). This cluster centred on the IFG pars orbitalis (IFG_{or}) and extended in the superior direction to encompass part of mid vlPFC, although this activation was more posterior to the region modulated by the main effect of representation. Fig. 4B indicates that the interaction was driven by an effect of salience when making a belief representation, where having a highly salient self perspective attracts greater resources on this region than when self perspective is less salient. Whilst there appears to be a trend towards the opposite effect when representing non-mental information, a quadrupled t-test confirmed that the effect of salience was only present during mental representation in vlPFC. Thus, no other contrasts within this quadrupled t-test, including physical high > physical low, survived cluster detection. The

significant t-test further confirmed different neural demands in high versus low salience when making a belief inference in a single cluster, which comprised left mid to anterior vlPFC (Fig. 27C; Fig. 27D. green shading). The cluster peak for belief high > belief low was positioned in IFG_{tr} extending in the caudal direction to encompass voxels identified within the factorial analysis for the effect of representation (Fig. 27D. yellow shading), and in the ventral direction to include voxels identified by the factorial interaction effect (Fig. 27D. cyan shading). Table 7 details contrasts performed in these analyses and the resulting cluster peaks.

Table 6 Confirmatory Whole Brain Analysis, where False Belief > False Photograph

Cluster Peak	Hemi	Brodmann Area	Cluster size (voxels)	MNI coordinates			Z-value
				x	y	z	
Temporoparietal junction	R	21	1204	62	-56	16	9.53
Precuneus	L/R	7	2441	0	-56	34	9.34
Temporoparietal junction	L	21	767	-54	-56	22	8.35
Temporal Pole	R	21	442	56	6	-34	8.26
Middle Temporal Gyrus, anterior division	L	20	318	-54	0	-36	8.25
Middle Temporal Gyrus, posterior division	R	20	82	54	-8	-20	6.97
Frontal Pole	L	9	59	-14	52	32	6.81
Frontal Pole	L/R	10	110	0	62	10	6.62
Cerebellum	L	N/A	33	-28	-80	-40	6.59
Middle Frontal Gyrus	R	9	20	28	26	34	6.44
Middle Temporal Gyrus, posterior division	L	21	40	-64	-22	-10	6.43
Middle Temporal Gyrus, posterior division	R	21	16	66	-26	-10	6.19
Temporal Pole	L	20	2441	-38	16	-42	5.68

Note. Result reflects a two sample paired t-test, were neural regions listed were more responsive to either (B) over Physical (P) representation. Thresholded voxelwise at $p_{corr} < 0.001$). Brodmann Areas are approximate.

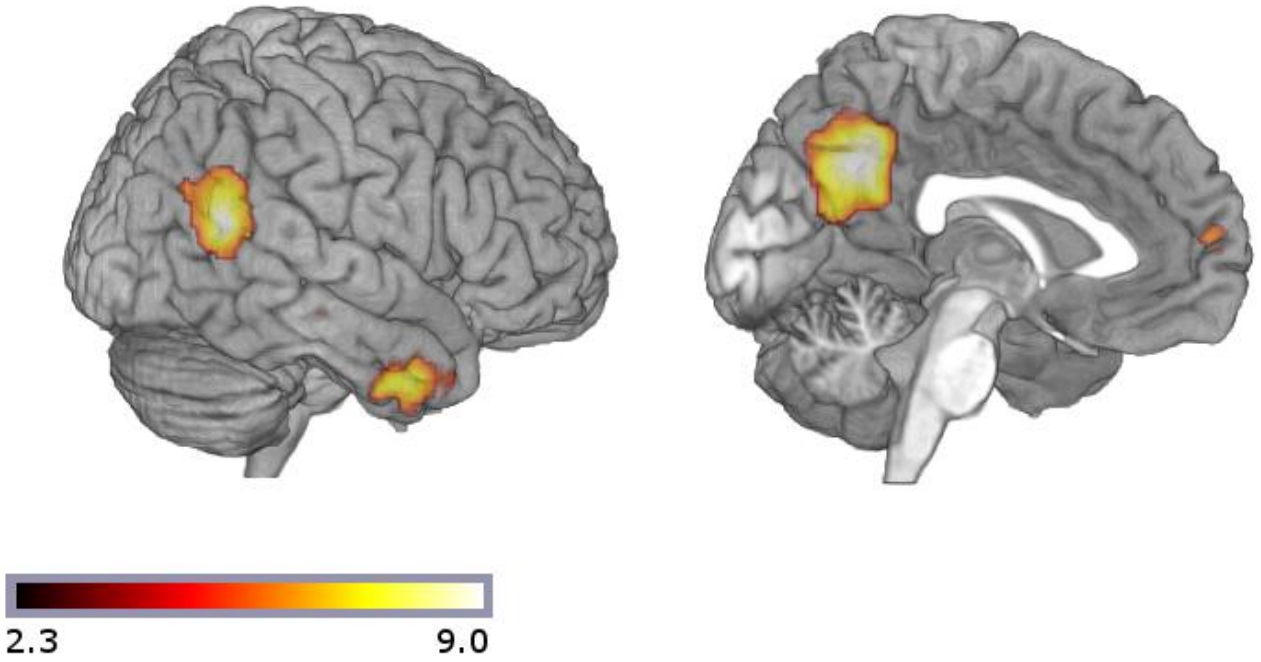

Fig. 26. Results: Modified ToM Localizer Confirmatory Analysis
Activation map for the contrast B > P overlaid onto the MNI brain template. Shows significantly activated voxels where $p_{corr} < 0.001$. Images reflect Z-corrected t-stat images. Lateral view shows right hemisphere to illustrate right TPJ.

Table 7 Cluster Peaks for Representation by Salience Analyses

Cluster Peak	Hemi	Brodmann Area	Cluster size (voxels)	MNI coordinates			Z-value
				x	y	z	
2x2 Repeated Measures ANOVA							
Main effect of Representation							
Inferior frontal gyrus, pars triangularis	L	45	605	-44	40	4	5.07
Inferior frontal gyrus, pars triangularis	R	45	280	48	40	10	4.11
Main effect of Salience of Self Perspective							
ns							
Representation * Salience Interaction							
Inferior frontal gyrus, pars orbitalis	L	47	246	-36	30	-10	4.49
Quadrupled t-test							
BH > BL							
Inferior frontal gyrus, pars triangularis	L	45	669	-44	38	10	4.38
BL > BH							
ns	-	-	-	-	-	-	-
PH > PL							
ns	-	-	-	-	-	-	-

PL > PH							
ns	-	-	-	-	-	-	-

Note. Results from ANOVA reflect regions identified using F-contrasts in a 2-way repeated measures factorial analysis with Representation (belief/physical, B/P) and Salience of Self Perspective (high/low, H/L) as within subjects factors. ANOVA results reflect cluster peaks for cortical regions which are modulated by varying representation status (belief/physical) and salience status (high/low) Quadrupled t-test reflects planned contrasts within specific conditions of interest. Brodmann Areas are approximate. ns = non-significant at $Z > 2.3$, $p_{corr} < 0.05$. B = belief, P = photo, H = high salience, L = low salience.

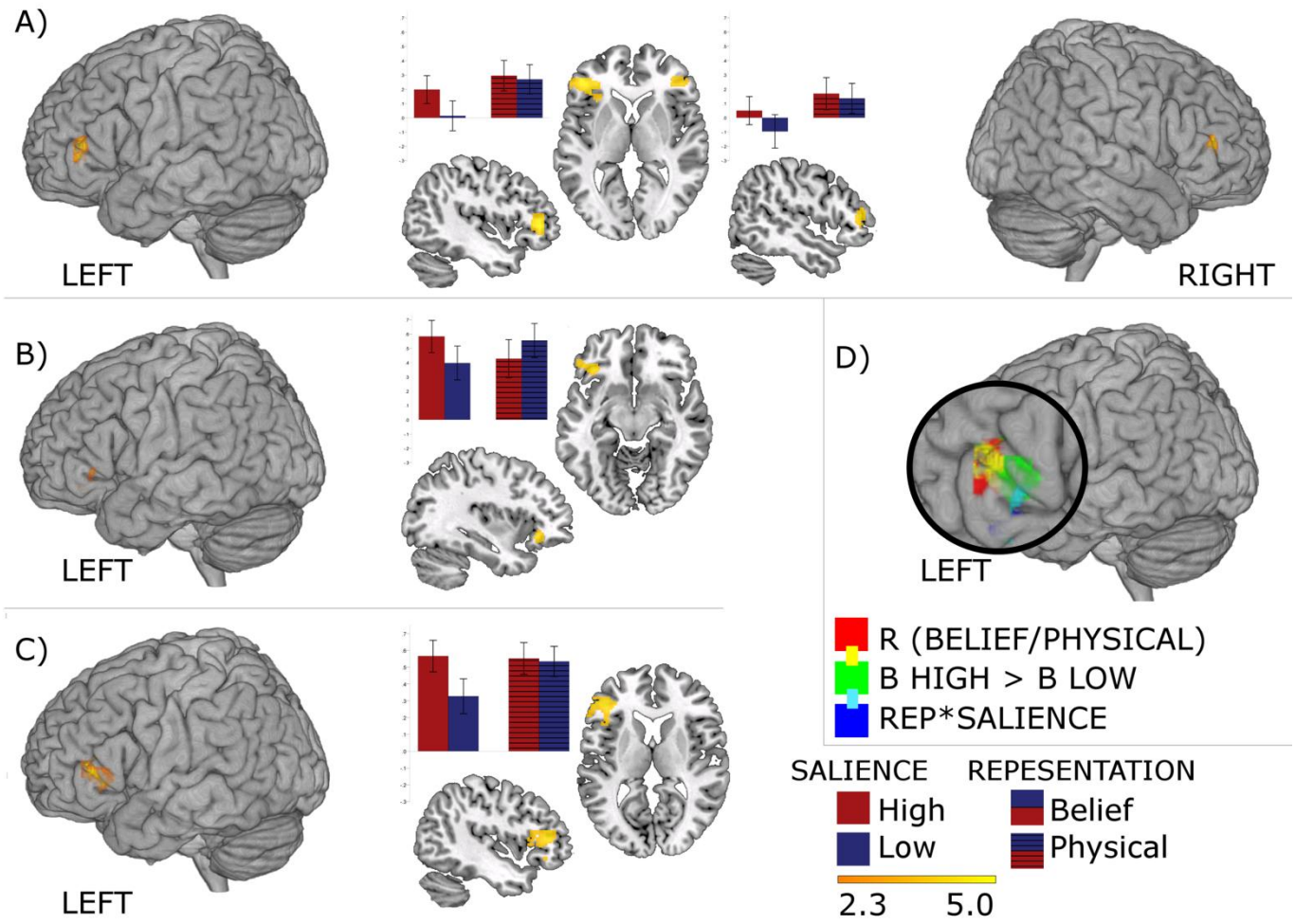

Fig. 27. Results: Modified ToM Localizer Factorial Analysis

Activation maps overlaid onto the MNI brain template show significantly activated voxels where Z > 2.3, $p_{corr} < 0.05$. Images are displayed in neurological convention, where left is represented on the left side of the image. Slices highlight cluster peaks. Plots reflect group mean

percentage signal change in the highlighted cluster for each condition; error bars reflect +/-1 SE of the mean. (**Panel A**) Result from 2x2 repeated measures ANOVA, with Representation (B/P) and Salience (H/L) as within-subjects factors. Image reflects a Z-corrected F-stat image of neural regions modulated by the factor of Representation (B/P). Slices from left to right, z = 4, x = -44, x = 48 (**Panel B**) Image reflects a Z-corrected F-stat image of regions modulated by an interaction between Representation (B/P) and Salience of Self Perspective (H/L). Slices z = -10, x = -36. (**Panel C**). Image reflects a Z-corrected t-stat image following a quadrupled t-test where BH > BL. Slices z = 10, z = -44. (**Panel D**) Activation maps are rendered onto a standard brain, showing the left vlPFC. Red shading indicates main factorial effect of Representation (B/P); Dark blue indicates Representation*Salience interaction; Green indicates significant result from quadrupled t-test where BH > BL. Yellow indicates voxels recruited by both factor of Representation (B/P) and quad t-test BH > BL; Cyan indicates voxels recruited by R*S interaction and BH > BL t-test.

DISCUSSION

The present study aimed to examine how vlPFC responds to variation in the salience of self perspective, in mental versus non-mental representation. Based on the behavioural literature, an effect of salience was tentatively anticipated in task performance, where interference from own perspective in high salience representation was expected to inflate error rates (Birch & Bloom, 2004; Birch & Bloom, 2007; Carlson & Moses, 2001; German & Hehman, 2006). This prediction was made with caution, due to the ease with which typical adults were expected to find this relatively slow-paced paradigm. On account of converging evidence that vlPFC plays an important part in ToM (Mar, 2011; Spreng et al., 2009), and drawing from neuropsychological (Samson et al., 2005), ERP (McCleery et al., 2011) and fMRI data (Rothmayr et al., 2011; van der Meer et al., 2011) suggesting that this role may reflect an inhibitory mechanism, variation in the salience of self perspective was expected to modulate vlPFC. Conversely, previously published results from the original localizer task suggest that vlPFC would not respond to variation in representational content, that is, when making a belief versus a physical representation.

In general, adult participants were skilled at representing both outdated mental and physical information that was at odds with reality. They demonstrated ease with mental and non-mental representation, as well understanding the real state of affairs. No effect of representation was identified in terms of the number of errors made; thus, adults showed no behavioural cost between assuming a belief, versus physical, representation of an event. This is in line with research which demonstrates that young children find similar difficulty with both types of representation (Zaitchik, 1990) and, when executive demands are tightly controlled, adults with brain damage perform equivalently across the two tasks, where deficit or success with false belief reasoning is accompanied by respective deficit or success in false

photograph reasoning (Apperly, Samson, Chiavarino, Bickerton, & Humphreys, 2007). A significant behavioural cost was, however, associated with making a representational inference under the influence of a highly salient self perspective. This is in line with numerous behavioural data which suggest that own perspective interferes when making judgments about a naive or misinformed other (see Birch & Bloom, 2004; 2007 for a review). For example, the classic unexpected transfer task illustrates that children under the age of four respond from their own, egocentric perspective (Wimmer & Perner, 1983). Furthermore, adults tend to overestimate the extent to which own knowledge is shared by others, in terms of the outcome of an event (Fischhoff, 2003) or general knowledge (Thomas & Jacoby, 2013). Also, adults, like children, suffer interference from their own visual perspective when considering the viewpoint of an agent (Keysar et al., 2003; Surtees & Apperly, 2012). Considered together, these data add further credence to the view that holding an incompatible perspective with an agent, when realised, is necessarily effortful.

A whole brain analysis, using the approach specified in Saxe and Kanwisher (2003), confirmed that the modifications presented here had not altered the operation of the localizer. As a result, TPJ and mPFC were identified when representing a false belief over a false photograph, or the like. Nonetheless, the novel factorial analysis identified that bilateral vlPFC was responsive to representational content. Whilst the original localizer does not detect this difference using the whole brain approach, it is worth highlighting that there is no contradiction between the result from the standard localizer contrast and the factorial analysis. The voxels identified by this factor are, overall, more positively activated for physical versus belief representation. As a consequence, one would not expect to see vlPFC in a whole brain analysis that computed belief > physical. Its inverse, however, should recruit this region. This assertion was confirmed post-hoc by computing physical > belief, which resulted in bilateral

vlPFC. In addition, the difference between mental and non-mental representation is amplified in this subset of voxels by the low salience belief condition. Importantly, the difference between types of representational content when self perspective is highly salient, as is the standard approach with these vignettes, is negligible. Furthermore, due to constraining the analyses to the frontal cortex, the factorial model had increased power over a whole brain approach to detect subtle effects within vlPFC. For all of these reasons, the original vignettes would be highly unlikely to replicate this result if the appropriate contrast were to be computed.

The interaction effect and quadrupled t-test identified a difference in neural demands in left anterior vlPFC for high versus low salience of self perspective in belief representation. A reduction in resource demands was associated with lowering the salience of one's own perspective of the real state of affairs during mental representation. No such gain, however, was identified in vlPFC when reasoning about outdated non-mental content, such as photographs, videos and paintings. To consider the data as a whole, a distinct pattern exists, where the amplitude of activation in vlPFC is greater for high versus low belief, as a consistent feature. This region did not, however, appear to differentiate between representational content when self perspective was highly salient, or between salience of self perspective when making a physical representation. Taken together, these data suggest that the processes occurring across representational tasks are non-equivalent.

A possible imbalance in executive demands between representing false beliefs and false photographs, specifically, has been proposed on theoretical and behavioural grounds (Callejas, Shulman, & Corbetta, 2011; Muller, Zelazo, & Imrisek, 2005; Perner et al., 2006; Perner & Leekam, 2008; Russell et al., 1999; Sabbagh, Moses, & Shiverick, 2006). For example, Callejas et al. (2011) demonstrated non-equivalent working memory demands and

linguistic structures between the original false belief and false photograph localizer vignettes. More generally, theoretical concerns have been raised regarding the 'falseness' of the false photograph task, in so much that a photograph cannot really be false (Perner & Leekam, 2008). Both false belief and false photograph vignettes appear to elicit confliction between the world as it is now, against how it was at the time the representation was captured. Nonetheless, what the false photograph condition captures is actually an accurate representation of the referent that it captured, as it was at that moment in time. A false belief on the other hand, such as Lorraine's belief that her money is on the side table, is inaccurate with regards to the referent on which her belief centres – *the money is actually in her daughter's purse* (see Perner & Leekam, 2008 for a more comprehensive discussion). Whilst the literature support the general assertion that not all representation tasks are equivalent, they do not address the difference found in left anterior vlPFC, as only the neural profile of low versus high salience belief reasoning was statistically divergent. Instead, prior arguments have evaluated the more typical high inhibition belief versus high inhibition photo tasks, where we find no difference. Whilst we have do not have the scope to accept or refute prior claims regarding non-equivalence in the present study, we add a specific dimension to the argument, in that we demonstrate that the processes needed to successfully manipulate competing perspectives are distinct, on the basis of representational content. We therefore suggest that, where two information streams are incongruent, an amount of executive control will be exerted to discard erroneous responses. This effect is seen in social, as well as non-social, tasks such as Go/No-Go and Stroop. In belief reasoning, and perhaps in wider social tasks (see Surtees & Apperly, 2012), however, self perspective creates a unique problem, in that it interferes with the ability to appreciate what a misinformed or naive agent will know (Birch & Bloom, 2004; Birch & Bloom, 2007). One basis for this difficulty has been suggested to

reflect the self referent serving as a default knowledge state (Leslie & Polizzi, 1998). When making sense of the world around us, we may use our own knowledge as a primary informant. This is likely to hold predictive value in situations where other social information is lacking. In all four experimental conditions, regardless of the salience manipulation, the self perspective and the representational content compete for attention. However, in the case of belief reasoning specifically, by reducing the salience of self perspective, the salience of the default state was reduced. This in turn, we propose, would reduce the amount of competition between the two information streams, therefore guiding attentional resources towards the representational content. Non-mental representation tasks, on the other hand, are unlikely to call upon the self referent in the first instance, as there is little predictive benefit in using self as a model for physical causality. Here then, the salience manipulation would be ineffectual in terms of reducing competition, as without a default reference state, information regarding reality and the representational content would carry equal weighting. Thus, we suggest that the processing streams for these tasks reflect different rules. These are set according to the nature of the representation, where reasoning about beliefs attracts semantic knowledge for ToM, including the self referent, whereas reasoning about photographs recruits semantic knowledge for artefacts (Apperly et al., 2007). On this basis, we propose that left IFG mediates the controlled retrieval of, and selection from, competing informational items, where competition between informational items reflects salience cues that are directed from semantic information stores.

In contrast to the current result, together, Vogeley et al. (2001), Samson et al. (2005) and McCleery et al. (2011), suggest a right lateralised process is involved in inhibiting competing perspectives. We put forward a refinement to this suggestion, in that the right vlPFC reflects the general process of suppressing irrelevant informational items. Note that, in

all conditions, two competing information streams were present; a representational and real state of affairs was always outlined. As a result, attention must always be directed away from one informational item. As a result, if IFG supports this general inhibitory mechanism, it would not be identified by the analyses that can be performed on the present data. We tentatively suggest, therefore, that the presence of right vlPFC may reflect a more generalised process of suppressing irrelevant distracters for the purpose of inhibition, whereas left vlPFC is involved in controlling the retrieval of competing informational items. Notably, this latter process can be facilitated in mental representation by reducing the salience of own perspective, as 'self' serves as a source of reference in situations of limited knowledge.

CHAPTER 6:

GENERAL DISCUSSION

INTRODUCTION

This thesis has presented four studies of adult Theory of Mind (ToM). Using techniques from cognitive neuroscience, the functional profile of the temporoparietal junction (TPJ) and medial prefrontal cortex (mPFC), considered core mentalizing regions, alongside ventrolateral prefrontal and dorsal medial prefrontal cortices (vlPFC/dmPFC respectively), regions more typically associated with executive processes, have been explored. Throughout, this thesis has attempted to bring together data from social, developmental and cognitive psychology with the methods pertained by neuroscience. The influence from developmental psychology should be apparent; thus, to have only collected data from typically functioning adult participants may seem somewhat strange. However, the approach used reflects the belief that working with a mature, fully functioning system is informative to understanding the development of ToM, as well as what might constitute or explain dysfunction (Apperly et al., 2009). The case in point is that every chapter in this thesis has demonstrated that the overall pattern of adult ToM reasoning is consistent with that of young children, albeit to a lesser magnitude; adults have been shown to be slower and more error prone when making judgements about an agent who holds a false belief or avoidance desire, for example. This, in itself, is revealing when working towards an account of ToM that explains when (contextually) and how (cognitively) executive processes are recruited. Whilst adults demonstrate greater proficiency with ToM than children, with careful experimental manipulation, it was possible to elicit behavioural and neural costs which might normally go undetected. For example, in Chapter 2, manipulation of the valence of an agent's ToM state highlighted a larger behavioural cost associated with belief versus desire. Likewise, the developmental literature has demonstrated that children typically are able to apply a simple desire based reasoning before they can use a belief based approach, but that negatively

valenced states attract the greatest difficulty (Rakoczy, Warneken, & Tomasello, 2007; Repacholi & Gopnik, 1997). This finding has already been replicated in adults elsewhere (Apperly et al., 2011; German & Hehman, 2006). The present thesis, however, provides new information regarding the basis of the increased behavioural cost associated more generally with belief over desire reasoning: whilst the valence of desire manipulation was shown to modulate neural executive selection resources in dmPFC, particularly dorsal ACC, the belief reasoning manipulation additionally modulated vlPFC, which is thought to serve an important role in inhibiting self perspective. Taken together, these data suggest that, in the adult brain, different neurocognitive processes are engaged when representing these specific belief and desire states. Consequently, though both belief and desire reasoning may draw on dmPFC, belief reasoning was shown to also engage vlPFC, due to incongruence in self versus other perspectives. This information, of course, leads to testable hypotheses regarding population sub-groups, such as children and clinical populations. Nevertheless, the manipulations presented here reflect the approach that has, more generally, been adopted in prior developmental research. It may be that systematic variation in congruence between self versus other preferences would also elicit activation in vlPFC. Thus, the appearance of greater difficulty with belief in the developmental literature is merely a reflection of previous paradigmatic biases. This question needs to be addressed in order to ascertain whether belief reasoning, in general, is simply more effortful than desire reasoning because there are additional processing costs associated with epistemic states, or whether this just appears so due to the nature of the questions previously asked. Moreover, such a question would be informative in terms of identifying the true nature of the conflict that has been identified in false versus true belief.

In these closing pages, a brief summary of each experiment and result is given. This is followed with sections which examine what the culmination of the studies presented in this thesis say about the role of TPJ, mPFC including dmPFC, and vlPFC in ToM. I close with concluding comments regarding how these regions may interact with each other, and how specific ToM contexts and concepts reflect the recruitment of these regions.

Summary of Empirical Results and General Conclusions

Chapter 2 described a belief-desire reasoning paradigm that was used to examine two theories of executive control in ToM. The first suggests that control is required in negatively valenced ToM states, such as false belief and avoidance desire, due to the need to disengage attention from one salient target in order to switch to another (Leslie & Polizzi, 1998). Based on neuroimaging studies of ToM and executive control, such a process was suggested to involve regions within dmPFC, such as the ACC and frontal eyefields (Rothmayr et al., 2011; Saxe, Schulz, et al., 2006; van der Meer et al., 2011). The key criterion that would evidence this process, however, was that executive control regions would be recruited due to systematic variation in valence. As a consequence, both factors of belief and desire would recruit this region. The second theory that was outlined proposes that false belief reasoning has the unique property over true belief and approach/avoidance desire, in that it generates incompatibility between the viewer's own perspective and that of the agent. As a result, additional control processes may be required to inhibit one's own privileged knowledge in false belief reasoning specifically (Birch & Bloom, 2004; Birch & Bloom, 2007). Based on neuroimaging (van der Meer et al., 2011; Vogeley et al., 2001), lesion (Samson et al., 2005) and event related potential (ERP) data (McCleery et al., 2011), this process was expected to recruit vlPFC. The paradigm outlined in Chapter 2 identified neural data which complemented

the predictions made by each of these theories. dmPFC appeared to generally index conflict resulting from varying the valence of each mental state. vlPFC responded due to variation in congruence between self and other perspectives, which was only a feature of false versus true belief reasoning. Together, these data suggest that ToM draws on different executive mechanisms, which are driven by the content of the ToM state that is being represented and, in special circumstances, how this relates to own perspective.

Chapter 2 also sought to provide clarity on how core elements of the ToM network, such as TPJ and mPFC, respond to fine-grained mental states. In addition to the belief-desire reasoning paradigm, the participants also completed a ToM localizer task (Saxe & Kanwisher, 2003). This enabled identification of the area of TPJ believed to be specific for ToM (Saxe & Kanwisher, 2003; Saxe & Powell, 2006; Saxe & Wexler, 2005; Scholz et al., 2009; Young, Dodell-Feder, et al., 2010), and scrutiny of signal changes which occurred as a result of the valence manipulation. These analyses determined that TPJ was modulated by variation in the content of mental states, where negatively valenced mental states drew more heavily on resources. This result suggests that the proposal that TPJ responds preferentially to transitory mental states (Van Overwalle, 2009) does not adequately describe the functional profile of this region. Instead, in Chapter 2, it was suggested that activation in TPJ may be regulated by the relative difficulty of each belief-desire state, where more complex mental states will draw more heavily on this region. Finally, Chapter 2 highlighted that rostral medial prefrontal cortex (rmPFC) was not recruited by the belief-desire task, despite this region being thought to be important for ToM (Amodio & Frith, 2006), and it having been identified in the same participants using the ToM localizer. In this context, Chapter 2 outlined the proposal that rmPFC facilitates mentalizing when one needs to make more elaborate judgments, such as when reasoning beyond the constraints of the given information. It was suggested that the

belief-desire reasoning task would not attract reasoning in such a manner, as participants were told precisely the agent's beliefs and wants. The localizer, on the other hand, invited a much deeper inference, where the agent's state of mind had to be deduced on the basis of a complex scenario.

The experiment in Chapter 3 sought to modulate rmPFC by systematically varying the mode of reasoning required to represent an agent's ToM. By making a minimal change to the belief-desire reasoning paradigm outlined in Chapter 2, an additional experimental level was added to the desire condition, wherein the agent's desire state was unspecified. This manipulation required participants to reason abductively about the agent's desire on the basis of sparse social information (a photograph). This new condition was included alongside the prior true/false belief and approach/avoidance desire states; all of which invited a more constrained, deductive ToM inference. The resultant experimental paradigm therefore invited both deductive reasoning, as featured in the original paradigm, and abductive reasoning. Importantly, a distinction was drawn between activation in dmPFC, which was argued to reflect conflict in Chapter 2, and anterior rmPFC, which is more typically described as a core part of the ToM network (e.g., Amodio & Frith, 2006). Consequently, it was proposed that, again, dmPFC would be recruited for both belief and desire reasoning, but that only reasoning about an agent who held an unspecified desire would recruit rmPFC. The data supported this assertion, where an area of anterior rmPFC was recruited preferentially when reasoning about an agent whose desire was unspecified. Thus, Chapter 3 demonstrated that activation in rmPFC can be explained by the mode of reasoning, where rich, abductive inferences recruit this region but constrained, deductive inferences do not.

Chapters 2 and 3 showed that vlPFC was recruited when the truth status of an agent's belief was manipulated, suggesting that this region was responsive to differences in the

perspective between self and other. As a result, Chapters 4 and 5 both focussed specifically on the role of vlPFC in inhibition of self perspective. The patient study outlined in Samson et al. (2005) provides the only causal evidence for the role of inferior frontal cortex in inhibition of self perspective. Patient WBA was shown to be impaired on classic, high inhibition false belief tasks. He was, however, highly capable at making ToM judgements when the salience of his own perspective was reduced. Vogeley et al. (2001) identified a small region of right vlPFC that was active when imagining self in a fictional scenario whilst simultaneously reasoning about an agent's ToM. Drawing on Vogeley's result, Samson et al. (2005) suggested that WBA's deficit may be attributable to damage in right vlPFC, which rendered him unable to overcome interference from his own perspective in the high salience tasks. Nevertheless, WBA's lesion was sizeable, making it difficult to have confidence regarding the precise localization of the cause of WBA's impairment.

In order to confirm a causal role for right vlPFC in supporting the inhibition of self perspective, a Transcranial Magnetic Stimulation (TMS) study was conducted. Using a modified version of the belief-desire task presented in Chapter 2, adult participants were required to identify the emotive outcome of an agent in a simple ToM scenario. Using an offline TMS protocol, data were collected from participants following the application of continuous theta burst stimulation to a coordinate set for vlPFC identified in Hartwright et al. (2012, [Chapter 2 here]). An interaction effect was predicted, where participants were expected to be slower and more error prone in false belief reasoning, but not true belief or desire reasoning, following TMS to right vlPFC versus a control site at the vertex. Despite replicating the overall behavioural effects of negatively valenced mental states, where these were shown to be more effortful, the effect of TMS was highly variable across participants. The effect on response latencies and error rates was heterogeneous, where almost half of the

participants showed a reduction in reaction time following TMS to vlPFC. A subgroup of these showed a further facilitatory effect, where an improvement in accuracy was identified. Later research has been published which demonstrates that the theta burst protocol, as was used in Chapter 4, is inefficacious in producing homogenous, consistent effects that are repeatable both within and across subjects (Hamada et al., 2012; McAllister et al., 2013), and may even produce a facilitatory drive from the contralateral region to the perturbed area (Hartwigsen et al., 2013). As a result, no firm conclusions were drawn in Chapter 4 regarding whether right vlPFC is casually involved in ToM scenarios that invite incongruence between perspectives.

The data presented in Chapters 2 and 3 suggested that vlPFC was recruited when the truth status of an agent's belief was manipulated. In Chapter 2, this was argued to reflect the fact that false, versus true, belief reasoning raises a discrepancy between own perspective of the real state of affairs and the agent's (misinformed) understanding of reality. Whilst compatible with a growing literature which implicates vlPFC in resisting interference from self perspective (Samson et al., 2005; van der Meer et al., 2011; Vogeley et al., 2001), it was important to make certain that this result was not, in fact, a reflection of modulation due to truth status. That is, that vlPFC contains a module which responds specifically to truth status and not to conflict between self and other perspectives. Whilst unlikely, it was important to disambiguate these possibilities. As a result, an alternative paradigm was devised in which the salience of the participant's perspective was manipulated within the context of a representational task, rather than using valence as before. By making a minor modification to the ToM localizer (Saxe & Kanwisher, 2003), the false belief and false photograph stimuli were modified so that the change of state that typifies these scenarios was either explicitly, or loosely, specified; thus resulting in either a highly, or minimally, salient self perspective

respectively. This experiment demonstrated that left vlPFC was responsive to salience, but only in the case of making a mental representation. For non-mental representation – the false photograph-type scenarios – left vlPFC did not differentiate between high or low salience of self perspective. These data led to the conclusion that left vlPFC is brought in to serve an underlying mechanism that is a necessary feature of selecting a perspective that is incompatible to one's own, such as resolving interference between the two perspectives, or inhibiting one's own, task-irrelevant viewpoint. On the basis of the left lateralised activation identified in Chapter 5, but bilateral activation in our other experimental protocols, right vlPFC was suggested to operate more generally in actioning the resolution of two competing information streams – a feature that was consistent across both levels within mental and non-mental representation conditions.

Activation in the Temporoparietal Junction Reflects ToM Content

Of the neural regions examined within this thesis, the TPJ, perhaps surprisingly, receives the least investigation. This in part reflects the already sizable literature which outlines its functional characteristics. It is also a reflection of how the research questions have naturally evolved in studying the neurocognitive basis of adult ToM. Nevertheless, the TPJ has been the cause of considerable debate, particularly regarding whether it reflects a domain-specific module for mentalizing, and thus responds in an undifferentiated manner towards ToM states.

The TPJ is identified with significant regularity in studies of ToM. Unlike mPFC, which is also regularly recruited for ToM, TPJ, however, reflects a more consistent profile, where it appears to respond preferentially to representing transient mental states like beliefs and desires, over control tasks, such as representing non-mental content like outdated

photographs or signs (see Van Overwalle, 2009 for a review). On this basis, a ToM localizer has been devised (Saxe & Kanwisher, 2003), where neural activation that is specific to inferring an agent's false belief over the contents of an outdated, as in 'false', photograph or sign, is identified. This activation is taken to reflect processes that are unique to ToM. This localizer demonstrates robust recruitment of TPJ, mPFC, precuneus and temporal poles (e.g., Aichhorn et al., 2009; Hartwright et al., 2012; Mitchell, 2007; Perner et al., 2006; Saxe & Powell, 2006; Saxe & Wexler, 2005; Scholz et al., 2009; Young, Dodell-Feder, et al., 2010). Using these locations for region of interest (ROI) analyses, further examinations into how each ROI responds to other socially relevant stimuli have been conducted. These studies suggest that, unlike the other neural regions recruited by the localizer, right TPJ, whilst preferentially engaged when reasoning about the mental state of an agent, does not discriminate between human and non-human physical descriptions, suggesting that it is not responsive simply to the mere presence of a human actor (Saxe & Kanwisher, 2003), and that it is not specifically engaged in response to descriptions regarding the social, cultural or geographical background of an agent, indicating that it is not just engaged by social content (Saxe & Wexler, 2005). Note that in individuals with autism, however, similar comparisons show that right TPJ does not differentiate between mental and non-mental representation (Lombardo, Chakrabarti, Bullmore, & Baron-Cohen, 2011)

Though these studies provide a convincing argument that the right TPJ may reflect a module for ToM, doubt has been raised on the basis of a parallel between the format of false belief vignettes, which typically follow the structure of the unexpected object transfer task (see Chapters 1 and 2), and the observation that right TPJ is recruited when attention is broken in order to reorientate towards task relevant stimuli (Mitchell, 2007) . A subsequent meta-analysis confirmed that false belief reasoning and attentional reorientation tasks jointly

recruit right TPJ (Decety & Lamm, 2007), calling into question the basis on which this region was responding for ToM. Saxe and colleagues, however, later demonstrated that a distinct subdivision exists in TPJ, where voxels associated with ToM were separate from those for attentional processes (Scholz et al., 2009; Young, Dodell-Feder, et al., 2010). Similarly, parcellation of this region has also been identified through structural imaging (Mars et al., 2012) and functional connectivity analysis (Bzdok et al., 2012). These later claims support the utility of the false belief false photograph contrast in matching executive demands, in terms of the general requirement to reorientate attention from one stimulus to another. This is an important point that I return to later, in relation to ToM processes and mPFC.

In Chapters 2 and 3, participants were required to predict the action of an agent on the basis of that agent's belief-desire state. In Chapter 2, the ToM localizer was used alongside the novel belief-desire reasoning paradigm to enable ROI analyses within TPJ for specific belief-desire states. Whole brain analyses in Chapters 2 and 3 determined that bilateral TPJ were modulated by the valence of an agent's ToM state; ROI analyses further indicated that this effect reflected a greater magnitude of activation for the negatively valenced mental states, false belief and avoidance desire. This result provided strong evidence that TPJ is not responsive to representing mental states per se, as this was a requirement across all conditions, but that its activity varies according to the valence of the mental state. On the basis that these negatively valenced mental states attracted the greatest processing costs, in terms of reaction time and accuracy data, we suggested that TPJ may be up- or down-regulated by the relative difficulty, as indexed by behavioural data, of each belief-desire condition (Hartwright et al., 2012). This supposition is in line with other neuroimaging studies which demonstrate greater activation in TPJ for false over true belief reasoning (Aichhorn et al., 2009; Sommer et al., 2007). Recently, Koster-Hale and Saxe (2013) put forward an

alternative, compatible suggestion, which similarly explains the pattern of both the behavioural and the neural data for each belief-desire state examined in Chapters 2 and 3. They suggest that TPJ exhibits predictive coding behaviour, where the magnitude of activity is related not to the value of the immediately perceived stimulus, but to the difference between the stimulus value and its predicted value. Such a neural code is likely to increase efficiency as more neural resources are devoted to new, and therefore unpredictable, stimuli. The authors propose that when behaviour is judged in the context of these minute by minute decisions, neural prediction is based on a complex generative model of thoughts and behaviours. Reduced neural responses reflect predicted behaviour; behaviours which match the predictive model. Thus, situations where smaller neural responses are identified, if reflecting a neural prediction code, will be associated with improved behavioural performance as a result of proficiency through familiarity.

The neural prediction model proposed by Koster-Hale and Saxe (2013) makes intuitive sense by suggesting that positively valenced mental states, true beliefs and approach desires, are likely to attract lower processing costs because their associated behaviours are more typical to what we experience in the social world. Accordingly, negatively valenced mental states attract greater processing costs as these reflect less predictable behaviours. Clearly, social agents act on the basis of what they believe to be true; thus, they may act in a way which is incompatible with our expectations due to alternative, and at times misinformed, perspectives on reality. This is likely to result in a neural prediction error which, in turn, results in increased activation in TPJ.

Whilst the paradigms outlined in Chapters 2 and 3 were not designed to identify regions specific for ToM, nevertheless, the use of the ToM localizer task in Chapter 2 provided some assurance that overlapping activations in TPJ reflected voxels which were

most likely to be preferentially activated for ToM, if such a module exists. As a result, the finding that activity in TPJ is directly affected by the content of mental states should be considered as if these data reflect the functional profile of 'ToM nodes'. Though the data here do not speak to the case of domain specificity in ToM, it is noteworthy that neuroimaging work with people who have autism identifies a different functional response in right TPJ to representational tasks, which distinguishes these from neuro-typical participants (Lombardo et al., 2011). However, on the basis that parts of TPJ are known to reflect attentional reorienting, and that this can be a confounding feature in ToM tasks (Mitchell, 2007), it may simply be the case that people with autism have different attentional biases, or that they have no bias at all towards social stimuli. Alternatively, there may be no general, higher-cognitive framework on which to base neural predictions. All of these possibilities provide interesting avenues for further research; nevertheless, such speculation cannot be resolved with the present data. Moreover, assurances are needed that the analyses in clinical studies take into account the known functional subdivisions within TPJ (Bzdok et al., 2012; Mars et al., 2012) as autism is also characterised by executive dysfunction (Hughes, Russell, & Robbins, 1994).

A Functional Subdivision within the Medial Prefrontal Cortex

The mPFC is regularly recruited in ToM studies (e.g., see reviews by Carrington & Bailey, 2009; Lieberman, 2007; Mar, 2011; Spreng et al., 2009; Van Overwalle, 2009). Whilst researchers have debated its centrality in ToM (cf. Frith & Frith, 2006; Saxe & Wexler, 2005), the frequency with which ToM tasks elicit activation in this region suggests that it is likely to play an important role in mentalizing. The experimental paradigms outlined in Chapters 2 and 3 sought to examine the functional profile of this region. This was with the intention of identifying specific processes which affect recruitment of mPFC and, in turn,

using these to highlight anatomical divisions in this large cortical area. The study featured in Chapter 2 comprised two separate experiments. The first varied the valence of an agent's belief and desire state, with a view to unpacking two theories of executive control in ToM. These theories predicted different neural profiles, which were explicable on the basis of the wider literature on executive functions. Behavioural data suggest that certain mental states attract greater processing costs; adult participants are slower and more error prone when applying a false versus a true belief, or an avoidance versus an approach desire, to an agent (Apperly et al., 2011; German & Hehman, 2006). Thus, drawing parallel with research into the neural basis of executive functions, such a manipulation was used to determine which frontal activations were consistent with the predictions made by the aforementioned theories. The second experiment, a ToM localizer (Saxe & Kanwisher, 2003), varied the content of representational information from mental to non-mental; a protocol which has been used extensively to localise neural regions which are specific to ToM. Though each of the two experimental approaches feature limitations if used singularly, together, they permit powerful inferences regarding important processes for ToM, which have often been obscured by more typical experimental protocols (Hartwright et al., 2012 [Chapter 2 here]).

Chapter 2 demonstrated that representing behaviourally effortful ToM states, such as false belief and avoidance desire, was associated with dissociable patterns of frontolateral and dorsomedial prefrontal activation. Whilst lateral activations are discussed separately in due course, for the purpose of discussion here, it is sufficient to state that only variation in the valence of an agent's belief state modulated the lateral frontal cortex. Manipulation of both belief and desire valence, however, attracted common modulation of dmPFC, particularly the dorsal ACC. Chapter 3, which adopts a structurally similar paradigm, replicated these results.

Leslie and colleagues posit that belief-desire reasoning is supported by a common process which switches attention in cognitively demanding situations, for example, when attributing negatively valenced mental states, such as false belief and avoidance desire, to an agent (Friedman & Leslie, 2004; Friedman & Leslie, 2005; Leslie et al., 2005; Leslie & Polizzi, 1998). By manipulating the valence of an agent's mental state, the analyses presented in Chapters 2 and 3 isolated those neural regions which support behaviourally effortful ToM states. This is important; such effort is associated with significant processing costs throughout the lifespan (e.g., Apperly et al., 2008; Apperly et al., 2011; Cassidy, 1998; German & Hehman, 2006; Wimmer & Perner, 1983), thus, identifying their brain bases highlights neural regions to investigate in the development of ToM, and in the case of its dysfunction. Consistent with Leslie and colleagues, and prior behavioural work with adults (Apperly et al., 2008; Apperly et al., 2011; German & Hehman, 2006), false belief and avoidance desire were shown to pose behavioural difficulty even in adults who should be proficient at mentalizing. This behavioural cost was associated with activation in dmPFC. Importantly, although no belief or desire inferences were required in the paradigm outlined in Chapter 2 (and in all bar one condition in Chapter 3), the need to attend to the mental state of an agent was present in all of the experimental conditions within the belief-desire reasoning paradigms. As a result, it does not follow that activation in dmPFC is due to the act of representing the mental state of an agent (i.e. applying a ToM). Similarly, the localizer task, which should identify neural regions that are specialised for representing mental, but not physical, content, fails to identify dmPFC in the same group of participants (Chapter 2). This localizer task has been designed so that it necessarily subtracts out executive processes which manage attention between 'false' and 'true' locations – a consistent feature in both false belief and false photograph conditions. Therefore, an absence of dmPFC is indicative that, either this region is recruited by both false

belief and false photograph reasoning, suggesting a domain general process that is requisite to both tasks, or that dmPFC is recruited by neither. On the basis of the result from the initial belief-desire reasoning paradigm, the former suggestion is more likely. Taken together, then, these data suggest that more dorsal activation of mPFC, particularly dorsal ACC, reflects domain general resources. This viewpoint converges with other studies of belief valence (Sommer et al., 2007) and those that have used separate executive control and ToM tasks to identify areas of commonality (Rothmayr et al., 2011; Saxe, Schulz, et al., 2006; van der Meer et al., 2011). On the basis of the belief-desire reasoning paradigm presented in Chapters 2 and 3, and from the localizer results which feature in Chapter 2, it is suggested that dmPFC, particularly ACC, does not constitute a module for ToM. Instead, this region is likely to be modulated by variation in attentional demands. Such variation occurs when one is required to hold in mind and switch attention between multiple informational items, as is the case in false belief and avoidance desire (Friedman & Leslie, 2004; Friedman & Leslie, 2005; Leslie et al., 2005; Leslie & Polizzi, 1998) . Thus, a general executive role is assigned to this region in ToM.

Chapter 3 outlined three processes – representation, control and reasoning – which could be used to explain how specific regions of mPFC are involved in mentalizing. The first process describes the basic tenet on which the widely used ToM localizer (Saxe & Kanwisher, 2003) is based; that is, that ToM specifically requires representation of mental state content. Whilst touched upon within the current thesis, this process was not a specific focus of any of the studies outlined here. Nonetheless, arguments for such a case are presented elsewhere (e.g., see Frith & Frith, 2006; Gallagher & Frith, 2003). The second process, control, has already featured in prior discussion which relates dmPFC to supporting attentionally demanding scenarios. Further control processes subserved by lateral prefrontal regions are

outlined towards the latter part of the current chapter. Lastly, then, it was proposed that different approaches to ToM reasoning activate alternative modes of inference. rmPFC is recruited by numerous, what appear to be highly diverse, circumstances such as mentalizing under uncertainty (Jenkins & Mitchell, 2009), autobiographical thinking (Schacter et al., 2012; Spreng et al., 2009), prospection, and even when at rest (Spreng et al., 2009). In both Chapters 2 and 3, the mode of reasoning employed during ToM influenced whether anterior rmPFC was recruited. Richer, abductive inferences, such as those required in the ToM localizer task (Chapters 2) recruited this region, whereas constrained, deductive inferences, such as those required in the initial belief-desire reasoning paradigm (Chapter 2), did not. Together, these findings indicate that rmPFC supports an inferential process of free thinking that is required in numerous aspects of social cognition. This process can broadly be described in terms of the need to reason abductively.

This theory was tested in Chapter 3 by including a novel, unspecified desire condition, alongside the existing, specified belief-desire conditions. This minimal change was sufficient to elicit substantial recruitment of rmPFC, which was previously absent when this condition was not included; thus, suggesting that this region is brought in to service rich, deep representational processes. There are, however, two challenges to this proposal. The first relates to suggestions that rmPFC is specific for stimulus *dependent* thought. For example, when comparing neural regions that are required for stimulus independent versus stimulus orientated thought, rmPFC is recruited when required to attend to the external environment without generating any information internally (Gilbert, Simons, Frith, & Burgess, 2006). This contradicts the proposed role for rmPFC in reasoning beyond the constraints of the stimuli. There are, however, good reasons for believing that the work of Gilbert and colleagues (2006) presents no challenge to the new theory outlined in this thesis. The recruitment of rmPFC in

Chapter 3 was extensive. Consequently, it is unlikely that the voxels within this result reflect a single function. The unspecified desire condition, unlike the clearly specified approach and avoidance desire conditions, is likely to have attracted greater attention towards the stimuli provided. Participants were explicitly requested to make an inference regarding the agent's desire state on the basis of the visual appearance of the agent in a photograph. As a result, it is quite likely that part of this activation in rmPFC can be explained by an attentional mechanism which focuses one towards the salient features of the stimuli. Nevertheless, later work by Gilbert et al., (2007) highlights a rostral-caudal division within rmPFC, where ToM recruits a distinct neural population to that which was ascribed to attentional control. Thus, the authors conclude that there is no contradiction between activation described in rmPFC for mentalizing, which we propose reflects reasoning beyond the constraints of the stimuli (Hartwright, Apperly, & Hansen, 2013, [Chapter 3 here]), and activation for attention to stimulus dependent thought, which may also have been captured in Chapter 3. Inspection of the cluster identified in the unspecified desire condition in Chapter 3, against the rostral-caudal division in rmPFC outlined in Gilbert et al. (2007), suggests that the abductive condition captured both of these processes. Thus, the data find support for both types of inference in rmPFC; on the one hand, a general attentional process that may be incidental to the task and, on the other, a rich inferential process which is likely to reflect ToM reasoning outside of the laboratory. It would be informative to experimentally determine this, for example, by conducting a further study where both elements are systematically varied within a single paradigm.

The second challenge to the proposal that rmPFC supports a rich, inferential process relates to the ToM localizer. Recruitment of rmPFC by the ToM localizer (e.g., see Chapters 2 and 5) was earlier stated as one of the observations on which this abductive inferential process

was formulated. Yet, on the surface, the protocol of using short vignettes in this task may predict that both false belief and false photograph stimuli would invite abductive reasoning. As a result, rmPFC would not be identified by a contrast which subtracted activation that was common to the two tasks. There are grounds, however, for thinking that the level of abstraction is not wholly equivalent in the two tasks. Chapter 5 provided some discussion regarding the different nature of processing that is likely to be attached to mental, versus non-mental, representation. This is also further discussed in the present chapter, a little later. Nevertheless, the core argument, in this context, centres on theoretical concerns regarding the 'falseness' of the false photograph task. In short, Perner and Leekam (2008) suggest that a photograph cannot truly be false. It is an accurate representation of the referent that was captured in that moment in time. When considered in relation to the proposed reasoning process, the possibilities associated with reasoning about the contents of a photograph are therefore, in themselves, relatively constrained. Reasoning about an agent's belief, on the other hand, can attract innumerable representational possibilities; what the agent was thinking then, what the agent is thinking now, how are they going to feel when they discover occurrence X, what they might do when they discover occurrence X, and so on. As such, then, I suggest that the ToM localizer contrast is skewed towards abductive reasoning. Thus, as well as recruiting regions which may be specialized for ToM itself, it is also likely to recruit regions which are more active when reasoning beyond the constraints of the information provided. Note that the two are not synonymous. As was demonstrated in Chapters 2, 3 and 4, it is entirely possible to apply a ToM without inferring a mental state. Regardless, neural recruitment from the localizer task (Chapters 2 and 5) and the unspecified desire reasoning task (Chapter 3), converge on a region of anterior rmPFC which is consistent with the area identified in Gilbert et al. (2007) for supporting rich, abductive inferences.

A Laterality Effect in the Ventrolateral Prefrontal Cortex

The vlPFC, including the left inferior frontal gyrus (IFG) in particular, has been described in quantitative meta-analyses of ToM as a likely candidate for inclusion in the core mentalizing network (Mar, 2011; Spreng et al., 2009), yet its functional profile has been little explored. Whilst a handful of researchers have speculated that vlPFC supports inhibition of self perspective (Samson et al., 2005; van der Meer et al., 2011), the nature of such a process was unclear. This thesis has detailed three experiments which carefully manipulated psychologically relevant parameters, such as belief valence and salience of self perspective, in order to vary the magnitude of incongruence between the perspective of self and other (Chapters 2, 4 and 5). This approach was expected to modulate vlPFC, with the overarching aim of describing the functional profile of this region in ToM.

Chapters 2 and 3 demonstrated that recruitment of bilateral vlPFC reflected whether making an action prediction was based on an agent whose belief state was positively or negatively valenced. The critical distinction here being that, in false versus true belief reasoning, the participant's own privileged knowledge of reality was in direct conflict with the agent's. Thus, own perspective would need to be set aside in order to assume the competing knowledge state of the agent. Chapter 2 further clarified that the activation of vlPFC was unlikely to be attributable to general attentional differences, as the valence of an agent's desire state, which causes no conflict between perspectives but features similar attentional demands, was unrelated to recruitment of vlPFC. Interestingly, however, note that in Chapter 3, the inclusion of a new unspecified desire condition, which required an abductive inference regarding an agent's desire, resulted in the recruitment of bilateral vlPFC. Though not the focus of Chapter 3, this result is indicative that the participants reflected on self perspective – what their own thoughts and desires were – when attributing a desire state to the

agent, when no other social information was available. This process was unlikely to be invited in the other desire reasoning conditions outlined in Chapters 2 and 3, as these were always clearly specified. Thus, this result is consistent with the view that vlPFC is responsive to interference from self perspective when assuming that of an agent.

The previously reviewed chapters have all indicated that vlPFC provides a mechanism with which to inhibit self perspective. On the basis of the data described in Chapter 5, however, the functional description of vlPFC was further refined. These data suggest that left vlPFC manages incongruence between self and other perspectives, in terms of controlling the retrieval of, and selection from, multiple competing informational streams. For example, in mental representation, left vlPFC was modulated by salience of self perspective. When making a ToM judgment and self perspective was highly salient, left vlPFC was recruited to a greater magnitude than when self perspective was minimally salient. Considerable behavioural data indicate that children and adults confer to their own knowledge as a marker for truth (Fischhoff, 2003; Roxβnagel, 2000; Thomas & Jacoby, 2013; Wimmer & Perner, 1983) suggesting that the self referent may act as a default state in certain scenarios (Leslie & Polizzi, 1998). Thus, if the prepotency of the default self referent is extinguished, or at least minimised, as was the case specifically in the low salience condition in Chapter 5, vlPFC does not need to work as hard to guide selection towards the appropriate informational content. When making a non-mental representation, however, no such default state can logically exist. Unlike thinking about the contents of another person's mind, which may sometimes be predicted using self as a model, no similar gain can be achieved by modelling the contents of a physical representation on the self referent. Thus, the process of representing false photographs is likely to be different from representing false beliefs, particularly in terms of demands on semantic knowledge (Apperly et al., 2007). Chapter 5 demonstrated that the

salience manipulation for non-mental representation did not affect demands on left vlPFC. Here, left vlPFC placed equal weighting on the real versus the representational content, regardless of how well specified the real state of affairs was. Thus, in the false photograph condition, competition was roughly equivalent between the two informational items for high and low salience presentation. Though this did not translate into a different behavioural pattern between representational states, note that the paradigm used in Chapter 5 was unlikely to be sensitive enough to detect very subtle differences in response accuracy. Regardless, the level of competition that left vlPFC reflects is predicted by the postulated processing streams adopted for mental versus non-mental representation.

When the need for self perspective inhibition was varied, the paradigms used in Chapters 2 and 3 saw bilateral activation of vlPFC, whereas this was not the case for the task outlined in Chapter 5. One possibility for the absence of right vlPFC in Chapter 5 is that it reflects the general process of suppressing irrelevant informational items. All of the experimental conditions in Chapter 5 required the participant to ignore their own knowledge of the real state of affairs, so that they could attend to the representational content. Consequently, if right vlPFC represents the mechanism which actions this inhibitory process, it would not be identified by the statistical contrasts computed in Chapter 5, as the requirement to inhibit a competing informational item is held constant across conditions. In Chapters 2 and 3, however, it is likely that the true belief condition attracts minimal inhibitory processes, as the informational content between self and other are congruent. As a consequence, I suggest that modulation of right vlPFC in Chapters 2 and 3 was reflecting variation in valence, where the false belief manipulation was driving activation in the right hemisphere, as a result of the need to inhibit the irrelevant self perspective. Left vlPFC, on the other hand, I propose is modulated by differential demands in controlling the retrieval and

selection of competing information streams. In the case of Chapters 2 and 3, this would reflect conflict due to incongruence between perspectives that was present in false, but not true, belief reasoning. Likewise, in Chapter 5, by reducing the salience of self perspective, interference was minimised in the belief condition from the competing, default perspective. In the high salience condition, however, competition will have remained high between two relevant targets. As a result, the high salience condition attracted greater demands on left vlPFC than the lower salience condition.

The proposed functional descriptions of vlPFC just outlined fits well with the existing literature on how lateral prefrontal regions are purported to support higher cognitive functions. For example, structural imaging data demonstrate that individual variation in inhibitory control is associated with white matter integrity in right vlPFC (Forstmann et al., 2008). More generally, lesions to right PFC result in performance errors which reflect a failure to adjust performance according to discrepant stimuli (Stuss & Alexander, 2007). Likewise, as has been referred to frequently throughout this thesis, patient WBA, who had an extensive right frontal lesion, was unable to perform ToM tasks which exposed him to a competing truth, thus requiring inhibition of salient informational content. He was, nonetheless, able to successfully negotiate ToM tasks where conflicting perspectives were avoided (Samson et al., 2005). The suggestion that right vlPFC is specifically required for inhibiting salient self knowledge was supported by van der Meer et al. (2011). Using a version of the Samson et al. (2005) paradigm that was modified for functional magnetic resonance imaging (fMRI), van der Meer et al. (2011) demonstrated that bilateral vlPFC was recruited for high versus low salience ToM tasks. Note the parallel, here, between the experimental manipulation in these two studies and the experiment outlined in Chapter 5. What is particularly noteworthy is that, despite very different paradigmatic approaches, the

results converge with the overall data and how these fit with the currently proposed function of vlPFC. Nevertheless, despite converging evidence that right vlPFC supports an important inhibitory process in ToM, the TMS study in Chapter 4 failed to identify a causal role for this region in reasoning about an agent whose perspective was incongruent with reality. There are, however, recently published data which suggest that the heterogeneous effect identified may be due to the particular TMS protocol that was used (Hamada et al., 2012; Hartwigsen et al., 2013; McAllister et al., 2013). A further perturbation study using an alternative TMS protocol may be fruitful, particularly in light of later evidence outlining where right vlPFC may feature in the time course of ToM (McCleery et al., 2011).

The literature in the domain of executive function suggests that right vlPFC plays a secondary role in supporting the left homolog, where right lateralised vlPFC guides inhibitory processes when necessary (Badre, Poldrack, Pare-Blagoev, Insler, & Wagner, 2005; O'Reilly, 2010). Left vlPFC, on the other hand, is argued to work in a top-down fashion to drive selection and retrieval (O'Reilly, 2010). Stuss and Alexander (2007) suggest that left vlPFC organises the schemata necessary to complete a set task. Badre and Wagner (2007) expand on this, suggesting that structural connectivity between left vlPFC and lateral temporal regions implicate left vlPFC in the controlled retrieval of semantic information. Earlier, the suggestion was made that the processing streams involved in representation are likely to reflect different rules according to the nature of the representation; reasoning about beliefs may attract semantic knowledge for ToM, whereas reasoning about photographs and the like would recruit semantic knowledge for artefacts (Apperly et al., 2007). A process was described that related semantic content to the method of prioritizing competing informational units; ToM was suggested to default to the self referent, whereas no such default was available for non-mental representation. The processing consequences of which were

detectable in left vlPFC. In line with Badre and Wagner (2007), ToM is regularly seen to recruit lateral semantic association areas, such as the temporal poles (Olson, Plotzker, & Ezzyat, 2007). As can be seen in Chapters 2 and 5, the temporal poles were recruited by the localizer contrast which subtracts activation for false photograph from false belief reasoning. Taken together, this provides a plausible framework for the role of left vlPFC in controlling the retrieval of competing informational items, where competition between informational items reflects salience cues that are directed from semantic information stores.

LIMITATIONS

The method of cross-subject averaging that is typically used in neuroimaging research necessitates a homogenous participant group, where possible. The studies presented within this thesis all utilised participants who met strict inclusion criteria, in order to achieve an appropriate sample (see Appendix 1). For instance, all participants were strongly right handed, native English speakers and they were required to demonstrate aptitude with the tasks behaviourally. Whilst this approach is likely to reduce noise within the data and, thus, increase the power to detect relevant experimental effects, the use of a narrow participant sample is not without its limitations. Of course, one cannot be certain that the results can be extended to those groups who were ineligible to participate. There are, for example, data which demonstrate behavioural differences between those who have a dominant hand and those who are mixed handed (e.g., Rose, Jasper, & Corser, 2012). The use of right handed participants in neuroimaging studies, however, dominate the literature in order that any results are generalisable to the wider population, which is predominantly right handed. Though it is less clear whether the participants who performed the tasks adequately (and were therefore invited to participate in the neuroimaging experiments) were, themselves, in some way a-

typical, the causes that underpin error-rate are less explicable than those associated with performing a task accurately. Nonetheless, there remains the question as to whether there would be differences in ToM activation in those who could, versus those who could not, do the tasks.

This thesis outlined experiments which mainly used fMRI. There are, nevertheless, other approaches which would add further to our understanding of the social cognitive processes which underlie ToM. Patient studies, for example, allow causal inferences regarding brain function, which is not possible with fMRI. Nonetheless, neuropsychological studies are limited by the availability of patients with similarly circumscribed lesions, who also share similar pharmacological regimens. The use of virtual lesion techniques such as TMS, on the other hand, whilst also having the power to attribute causality, allow precise targeting of neural regions in neurologically intact participants. Unlike patient studies, virtual ablation permits one to obtain a homogenous sample which, as discussed previously, is generally desirable. Chapter 4 outlined a TMS experiment which sought causal evidence for the role of vlPFC in ToM. Although the results of the experiment in Chapter 4 were inconclusive, advancements that have occurred since conducting that particular study suggest new protocols which would disambiguate the previous null result. These protocols, therefore, warrant further investigation.

Though fMRI provides a powerful tool for localizing cognitive processes, the addition of structural imaging data, such as fibre tract information from Diffusion Tensor Imaging (DTI), would further add to our understanding of the social cognitive neuroscience of ToM. For example, performance in certain ToM tasks has been shown to correlate with white matter integrity which, in turn, has been shown to decline with age (Charlton, Barrick, Markus, & Morris, 2009). This result is perhaps unsurprising; nonetheless, questions regarding the ease

at which individuals are able to process mental state information may be explained not only by the activation of a network of neural regions, but by the integrity of the connections across that network. If we return to the exclusionary criteria for the tasks presented in this thesis, the basis of poor performance could, in part, be explained by differences in neural architecture. The use of DTI or other structural imaging methods would permit the investigation of questions such as this.

Understanding individual differences in ToM, both behaviourally and neurally, is an important next step from the work I have presented in this thesis. The additional approaches I have suggested in the preceding paragraphs provide possible ways of attempting this. Each experiment presented in this thesis was carefully constructed in order to address specific questions of interest. Importantly, these questions were underpinned by the belief that having a good understanding of the cognitive and neural bases of a typically functioning, adult ToM leads to testable hypotheses in developmental and clinical populations. The results from the experiments presented in this thesis promote further examination of a wider scope of questions, in terms of individual differences, the developing ToM, or atypical mentalizing, which in turn, can speak to some of the limitations addressed.

CONCLUSIONS

In presenting four studies of adult ToM, this thesis has described how experimental manipulation of control and reasoning processes can modulate activation in executive control and social brain regions. The tasks that have been outlined all manipulated psychologically relevant parameters within ToM tasks in order to modulate hypothesised neural regions in a pre-specified way. This approach identified the neural bases which underpin behavioural costs associated with representing specific ToM states in adult participants, some of which pose

challenge throughout the lifespan. Identifying the neural substrates of such costs, alongside their behavioural signatures, therefore, provides a useful evidence base on which to examine a-typical and neurodevelopmental functioning in ToM.

I have suggested that TPJ is not responsive to the content of mental states as such, but instead reflects neural effort applied to assimilate mental states which are more or less cognitively demanding. This suggestion is compatible with the predictive coding framework, outlined by Koster-Hale and Saxe (2013). I have evidenced a division in mPFC, where dmPFC, in particular dorsal ACC, reflects a domain general resource which is required when one needs to switch attention from one target to another. More rostral areas of mPFC, however, I propose are engaged when the ToM context invites rich, abductive inferences. Lastly, I suggest that vlPFC is involved in supporting inhibitory processes, when the ToM concept attracts conflict between salient perspectives. Considered as a whole, the neurocognitive data I have collected demonstrate how the social brain interacts with neural regions for executive function to facilitate a working ToM. Importantly, engagement of these executive regions is dependent on the ToM situation, as would be the case outside of the laboratory; the recruitment of executive regions for ToM is not simply due to incidental features such as poor experimental task design. The ToM network should, therefore, be thought of as a collection of brain regions which are flexibly engaged in order to adapt to the specific demands of the social world at that time.

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. doi: 10.1080/17470910903166853

Abu-Akel, A., & Shamay-Tsoory, S. (2011). Neuroanatomical and neurochemical bases of theory of mind. *Neuropsychologia, 49*(11), 2971-2984. doi: 10.1016/j.neuropsychologia.2011.07.012

Agnew, Z. K., Bhakoo, K. K., & Puri, B. K. (2007). The human mirror system: A motor resonance theory of mind-reading. *Brain Research Reviews, 54*(2), 286-293. doi: http://dx.doi.org/10.1016/j.brainresrev.2007.04.003

Aichhorn, M., Perner, J., Weiss, B., Kronbichler, M., Staffen, W., & Ladurner, G. (2009). Temporo-parietal junction activity in theory-of-mind tasks: falseness, beliefs, or attention. *Journal of Cognitive Neuroscience, 21*(6), 1179-1192. doi: 10.1162/jocn.2009.21082

Alaerts, K., Senot, P., Swinnen, S. P., Craighero, L., Wenderoth, N., & Fadiga, L. (2010). Force requirements of observed object lifting are encoded by the observer's motor system: a TMS study. *European Journal of Neuroscience, 31*(6), 1144-1153. doi: 10.1111/j.1460-9568.2010.07124.x

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*(4), 268-277. doi: 10.1038/nrn1884

Apperly, I. (2011). *Mindreaders: The Cognitive Basis of Theory of Mind*. Hove: Psychology Press.

Apperly, I. A. (2012). What is "theory of mind"? Concepts, cognitive processes and individual differences. *Quarterly Journal of Experimental Psychology, 65*(5), 825-839. doi: 10.1080/17470218.2012.676055

Apperly, I. A., Back, E., Samson, D., & France, L. (2008). The cost of thinking about false beliefs: Evidence from adults' performance on a non-inferential theory of mind task. *Cognition, 106*(3), 1093-1108. doi: 10.1016/j.cognition.2007.05.005

Apperly, I. A., Samson, D., Chiavarino, C., Bickerton, W.-L., & Humphreys, G. W. (2007). Testing the domain-specificity of a theory of mind deficit in brain-injured patients: Evidence for consistent performance on non-verbal, "reality-unknown" false belief and false photograph tasks. *Cognition, 103*(2), 300-321. doi: 10.1016/j.cognition.2006.04.012

Apperly, I. A., Samson, D., & Humphreys, G. W. (2005). Domain-specificity and theory of mind: evaluating neuropsychological evidence. *Trends in Cognitive Sciences, 9*(12), 572-577. doi: 10.1016/j.tics.2005.10.004

Apperly, I. A., Samson, D., & Humphreys, G. W. (2009). Studies of adults can inform accounts of theory of mind development. *Developmental Psychology, 45*(1), 190-201. doi: 10.1037/a0014098

Apperly, I. A., Warren, F., Andrews, B. J., Grant, J., & Todd, S. (2011). Developmental Continuity in Theory of Mind: Speed and Accuracy of Belief-Desire Reasoning in Children and Adults. *Child Development, 82*(5), 1691-1703. doi: 10.1111/j.1467-8624.2011.01635.x

Aron, A. R., Fletcher, P. C., Bullmore, E. T., Sahakian, B. J., & Robbins, T. W. (2003). Stop-signal inhibition disrupted by damage to right inferior frontal gyrus in humans. *Nature Neuroscience, 6*(2), 115-116. doi: 10.1038/nn1003

Aron, A. R., Robbins, T. W., & Poldrack, R. A. (2004). Inhibition and the right inferior frontal cortex. *Trends in Cognitive Sciences, 8*(4), 170-177. doi: 10.1016/j.tics.2004.02.010

Badre, D., Poldrack, R. A., Pare-Blagoev, E. J., Insler, R. Z., & Wagner, A. D. (2005). Dissociable controlled retrieval and generalized selection mechanisms in ventrolateral prefrontal cortex. *Neuron, 47*(6), 907-918. doi: 10.1016/j.neuron.2005.07.023

Badre, D., & Wagner, A. D. (2007). Left ventrolateral prefrontal cortex and the cognitive control of memory. *Neuropsychologia, 45*(13), 2883-2901. doi: 10.1016/j.neuropsychologia.2007.06.015

Balconi, M., & Canavesio, Y. (2013). high-frequency rtms improves facial mimicry and detection responses in an empathic emotional task. *Neuroscience, 236*, 12-20. doi: 10.1016/j.neuroscience.2012.12.059

Bargh, J. A., Chen, M., & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology, 71*(2), 230-244. doi: 10.1037//0022-3514.71.2.230

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a "theory of mind" ? *Cognition, 21*(1), 37-46. doi: http://dx.doi.org/10.1016/0010-0277(85)90022-8

Becchio, C., Cavallo, A., Begliomini, C., Sartori, L., Feltrin, G., & Castiello, U. (2012). Social grasping: From mirroring to mentalizing. *NeuroImage, 61*(1), 240-248. doi: http://dx.doi.org/10.1016/j.neuroimage.2012.03.013

Bernstein, D. M., Atance, C., Loftus, G. R., & Meltzoff, A. (2004). We saw it all along: visual hindsight bias in children and adults. *Psychological Science, 15*(4), 264-267. doi: 10.1111/j.0963-7214.2004.00663.x

Birch, S. A., & Bloom, P. (2004). Understanding children's and adults' limitations in mental state reasoning. *Trends in Cognitive Sciences, 8*(6), 255-260. doi: 10.1016/j.tics.2004.04.011

Birch, S. A., & Bloom, P. (2007). The curse of knowledge in reasoning about false beliefs. *Psychological Science, 18*(5), 382-386. doi: 10.1111/j.1467-9280.2007.01909.x

Botvinick, M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: an update. *Trends in Cognitive Sciences, 8*(12), 539-546. doi: 10.1016/j.tics.2004.10.003

Botvinick, M., Nystrom, L. E., Fissell, K., Carter, C. S., & Cohen, J. D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature, 402*(6758), 179-181. doi: 10.1038/46035

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*(1528), 2359-2367. doi: 10.1098/rstb.2009.0066

Brown, T. T., & Jernigan, T. L. (2012). Brain Development During the Preschool Years. *Neuropsychology Review, 22*(4), 313-333. doi: 10.1007/s11065-012-9214-1

Brune, M. (2005). "Theory of mind" in schizophrenia: A review of the literature. *Schizophrenia Bulletin, 31*(1), 21-42. doi: 10.1093/schbul/sbi002

Bush, G., Luu, P., & Posner, M. I. (2000). Cognitive and emotional influences in anterior cingulate cortex. *Trends in Cognitive Sciences, 4*(6), 215-222.

Bzdok, D., Schilbach, L., Vogeley, K., Schneider, K., Laird, A. R., Langner, R., & Eickhoff, S. B. (2012). Parsing the neural correlates of moral cognition: ALE meta-analysis on morality, theory of mind, and empathy. *Brain Structure & Function, 217*(4), 783-796. doi: 10.1007/s00429-012-0380-y

Call, J., & Tomasello, M. (2008). Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Sciences, 12*(5), 187-192. doi: 10.1016/j.tics.2008.02.010

Callejas, A., Shulman, G. L., & Corbetta, M. (2011). False Belief vs. False Photographs: A Test of Theory of Mind or Working Memory? *Frontiers in Psychology, 2*. doi: 10.3389/fpsyg.2011.00316

Carlson, S. M., & Moses, L. J. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development, 72*(4), 1032-1053.

Carlson, S. M., Moses, L. J., & Breton, C. (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*(2), 73-92. doi: 10.1002/icd.298

Carlson, S. M., Moses, L. J., & Hix, H. R. (1998). The role of inhibitory processes in young children's difficulties with deception and false belief. *Child Development, 69*(3), 672-691.

Carrington, S. J., & Bailey, A. J. (2009). Are there theory of mind regions in the brain? A review of the neuroimaging literature. *Human Brain Mapping, 30*(8), 2313-2335. doi: 10.1002/hbm.20671

Carter, C. S., Braver, T. S., Barch, D. M., Botvinick, M., Noll, D., & Cohen, J. D. (1998). Anterior cingulate cortex, error detection, and the online monitoring of performance. *Science, 280*(5364), 747-749.

Casile, A. (2013). Mirror neurons (and beyond) in the macaque brain: An overview of 20 years of research. *Neuroscience Letters, 540*, 3-14. doi: 10.1016/j.neulet.2012.11.003

Cassidy, K. W. (1998). Three- and four-year-old children's ability to use desire- and belief-based reasoning. *Cognition, 66*(1), B1-11.

Chambers, C. D., Bellgrove, M. A., Gould, I. C., English, T., Garavan, H., McNaught, E., . . . Mattingley, J. B. (2007). Dissociable mechanisms of cognitive control in prefrontal and premotor cortex. *Journal of Neurophysiology, 98*(6), 3638-3647. doi: 10.1152/jn.00685.2007

Charlton, R. A., Barrick, T. R., Markus, H. S., & Morris, R. G. (2009). Theory of mind associations with other cognitive functions and brain imaging in normal aging. *Psychol Aging, 24*(2), 338-348. doi: 10.1037/a0015225

de Graaf, T. A., & Sack, A. T. (2011). Null results in TMS: from absence of evidence to evidence of absence. *Neuroscience and Biobehavioral Reviews, 35*(3), 871-877. doi: 10.1016/j.neubiorev.2010.10.006

Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition. *Neuroscientist, 13*(6), 580-593. doi: 10.1177/1073858407304654

Dennett, D. C. (1978). Beliefs about beliefs. *Behavioral and Brain Sciences, 1*(04), 568-570. doi: 10.1017/S0140525X00076664

Devinsky, O., Morrell, M. J., & Vogt, B. A. (1995). Contributions of anterior cingulate cortex to behaviour. *Brain, 118 (Pt 1)*, 279-306.

Devlin, J. T., & Poldrack, R. A. (2007). In praise of tedious anatomy. *NeuroImage, 37*(4), 1033-1041. doi: 10.1016/j.neuroimage.2006.09.055

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. doi: 10.1016/j.neuroimage.2012.01.073

Dolan, M., & Fullam, R. (2004). Theory of mind and mentalizing ability in antisocial personality disorders with and without psychopathy. *Psychological Medicine, 34*(6), 1093-1102. doi: 10.1017/s0033291704002028

Duncan, J., & Owen, A. M. (2000). Common regions of the human frontal lobe recruited by diverse cognitive demands. *Trends in Neurosciences, 23*(10), 475-483. doi: 10.1016/s0166-2236(00)01633-7

Epley, N., Morewedge, C. K., & Keysar, B. (2004). Perspective taking in children and adults: Equivalent egocentrism but differential correction. *Journal of Experimental Social Psychology, 40*(6), 760-768. doi: 10.1016/j.jesp.2004.02.002

Eslinger, P. J., Grattan, L. M., Damasio, H., & Damasio, A. R. (1992). Developmental consequences of childhood frontal-lobe damage. *Archives of Neurology, 49*(7), 764-769.

Feredoes, E., Tononi, G., & Postle, B. R. (2006). Direct evidence for a prefrontal contribution to the control of proactive interference in verbal working memory. *Proceedings of the National Academy of Sciences of the United States of America, 103*(51), 19530-19534. doi: 10.1073/pnas.0604509103

Fischhoff, B. (2003). Hindsight not equal to foresight: the effect of outcome knowledge on judgment under uncertainty. 1975. *Quality & Safety in Health Care, 12*(4), 304-311; discussion 311-302.

Fodor, J. A. (1992). A theory of the childs theory of mind. *Cognition, 44*(3), 283-296. doi: 10.1016/0010-0277(92)90004-2

Forstmann, B. U., Jahfari, S., Scholte, H. S., Wolfensteller, U., van den Wildenberg, W. P. M., & Ridderinkhof, K. R. (2008). Function and structure of the right inferior frontal cortex predict individual differences in response inhibition: A model-based approach. *Journal of Neuroscience, 28*(39), 9790-9796. doi: 10.1523/jneurosci.1465-08.2008

Friedman, O., & Leslie, A. (2004). A developmental shift in processes underlying successful belief-desire reasoning. *Cognitive Science, 28*(6), 963-977. doi: 10.1016/j.cogsci.2004.07.001

Friedman, O., & Leslie, A. M. (2005). Processing demands in belief-desire reasoning: inhibition or general difficulty? *Developmental Science, 8*(3), 218-225. doi: 10.1111/j.1467-7687.2005.00410.x

Friston, K. J., & Henson, R. N. (2006). Commentary on: Divide and conquer; a defence of functional localisers. *NeuroImage, 30*(4), 1097-1099. doi: 10.1016/j.neuroimage.2006.02.007

Frith, C. D., & Frith, U. (2006). The Neural Basis of Mentalizing. *Neuron, 50*(4), 531-534. doi: 10.1016/j.neuron.2006.05.001

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society B: Biological Sciences, 358*(1431), 459-473. doi: 10.1098/rstb.2002.1218

Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of 'theory of mind'. *Trends in Cognitive Sciences, 7*(2), 77-83.

Gallese, V. (2007). Before and below 'theory of mind': embodied simulation and the neural correlates of social cognition. *Philosophical Transactions of the Royal Society B-Biological Sciences, 362*(1480), 659-669. doi: 10.1098/rstb.2006.2002

Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2*(12), 493-501. doi: 10.1016/s1364-6613(98)01262-5

Garavan, H., Ross, T. J., & Stein, E. A. (1999). Right hemispheric dominance of inhibitory control: An event-related functional MRI study. *Proceedings of the National Academy of Sciences of the United States of America, 96*(14), 8301-8306. doi: 10.1073/pnas.96.14.8301

German, T., & Hehman, J. (2006). Representational and executive selection resources in 'theory of mind': Evidence from compromised belief-desire reasoning in old age. *Cognition, 101*(1), 129-152. doi: 10.1016/j.cognition.2005.05.007

Gilbert, S. J., Simons, J. S., Frith, C. D., & Burgess, P. W. (2006). Performance-related activity in medial rostral prefrontal cortex (area 10) during low-demand tasks. *Journal of Experimental Psychology. Human Perception and Performance, 32*(1), 45-58. doi: 10.1037/0096-1523.32.1.45

Gilbert, S. J., Spengler, S., Simons, J. S., Steele, J. D., Lawrie, S. M., Frith, C. D., & Burgess, P. W. (2006). Functional specialization within rostral prefrontal cortex (area 10): a meta-analysis. *Journal of Cognitive Neuroscience, 18*(6), 932-948. doi: 10.1162/jocn.2006.18.6.932

Gilbert, S. J., Williamson, I. D. M., Dumontheil, I., Simons, J. S., Frith, C. D., & Burgess, P. W. (2007). Distinct regions of medial rostral prefrontal cortex supporting social and nonsocial functions. *Social Cognitive and Affective Neuroscience, 2*(3), 217-226. doi: 10.1093/scan/nsm014

Gopnik, A., & Astington, J. W. (1988). Childrens understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Development, 59*(1), 26-37. doi: 10.2307/1130386

Gough, P. M., Nobre, A. C., & Devlin, J. T. (2005). Dissociating linguistic processes in the left inferior frontal cortex with transcranial magnetic stimulation. *Journal of Neuroscience, 25*(35), 8010-8016. doi: 10.1523/jneurosci.2307-05.2005

Greenwald, A. G., & Banaji, M. R. (1995). Implicit social cognition - attitudes, self-esteem, and stereotypes. *Psychological Review, 102*(1), 4-27. doi: 10.1037//0033-295x.102.1.4

Hamada, M., Murase, N., Hasan, A., Balaratnam, M., & Rothwell, J. C. (2012). The Role of Interneuron Networks in Driving Human Motor Cortical Plasticity. *Cerebral Cortex, 23*(7), 1593-1605. doi: 10.1093/cercor/bhs147

Hamilton, A. F. (2013a). The mirror neuron system contributes to social responding. *Cortex*. doi: 10.1016/j.cortex.2013.08.012

Hamilton, A. F. D. (2013b). Reflecting on the mirror neuron system in autism: A systematic review of current theories. *Developmental Cognitive Neuroscience, 3*, 91-105. doi: 10.1016/j.dcn.2012.09.008

Hartwigsen, G., Saur, D., Price, C. J., Ulmer, S., Baumgaertner, A., & Siebner, H. R. (2013). Perturbation of the left inferior frontal gyrus triggers adaptive plasticity in the right homologous area during speech production. *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.1310190110

Hartwright, C. E., Apperly, I. A., & Hansen, P. C. (2012). Multiple roles for executive control in belief-desire reasoning: Distinct neural networks are recruited for self perspective inhibition and complexity of reasoning. *NeuroImage, 61*(4), 921-930. doi: 10.1016/j.neuroimage.2012.03.012

Hartwright, C. E., Apperly, I. A., & Hansen, P. C. (2013). Representation, Control, or Reasoning? Distinct Functions for Theory of Mind within the Medial Prefrontal Cortex. *Journal of Cognitive Neuroscience*. doi: 10.1162/jocn_a_00520

Hétu, S., Taschereau-Dumouchel, V., & Jackson, P. L. (2012). Stimulating the brain to study social interactions and empathy. *Brain Stimulation, 5*(2), 95-102. doi: http://dx.doi.org/10.1016/j.brs.2012.03.005

Hong, K. S., Lee, S. K., Kim, J. Y., Kim, K. K., & Nam, H. (2000). Visual working memory revealed by repetitive transcranial magnetic stimulation. *Journal of the Neurological Sciences, 181*(1-2), 50-55. doi: 10.1016/s0022-510x(00)00412-3

Hooker, C. I., Verosky, S. C., Germine, L. T., Knight, R. T., & D'Esposito, M. (2008). Mentalizing about emotion and its relationship to empathy. *Social Cognitive and Affective Neuroscience, 3*(3), 204-217. doi: 10.1093/scan/nsn019

Huang, Y.-Z., Edwards, M. J., Rounis, E., Bhatia, K. P., & Rothwell, J. C. (2005). Theta Burst Stimulation of the Human Motor Cortex. *Neuron, 45*(2), 201-206. doi: 10.1016/j.neuron.2004.12.033

Hughes, C. (1998). Executive function in preschoolers: Links with theory of mind and verbal ability. *British Journal of Developmental Psychology, 16*, 233-253.

Hughes, C., Russell, J., & Robbins, T. W. (1994). Evidence for executive dysfunction in autism. *Neuropsychologia, 32*(4), 477-492. doi: http://dx.doi.org/10.1016/0028-3932(94)90092-2

Iacoboni, M., & Dapretto, M. (2006). The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience, 7*(12), 942-951. doi: 10.1038/nrn2024

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *Plos Biology, 3*(3), 529-535. doi: 10.1371/journal.pbio.0030079

Jahanshahi, M., & Rothwell, J. (2000). Transcranial magnetic stimulation studies of cognition: an emerging field. *Experimental Brain Research, 131*(1), 1-9. doi: 10.1007/s002219900224

Jenkins, A. C., & Mitchell, J. P. (2009). Mentalizing under Uncertainty: Dissociated Neural Responses to Ambiguous and Unambiguous Mental State Inferences. *Cerebral Cortex, 20*(2), 404-410. doi: 10.1093/cercor/bhp109

Kalbe, E., Schlegel, M., Sack, A. T., Nowak, D. A., Dafotakis, M., Bangard, C., . . . Kessler, J. (2010). Dissociating cognitive from affective theory of mind: A TMS study. *Cortex, 46*(6), 769-780. doi: 10.1016/j.cortex.2009.07.010

Kerr, N., Dunbar, R. I. M., & Bentall, R. P. (2003). Theory of mind deficits in bipolar affective disorder. *Journal of Affective Disorders, 73*(3), 253-259. doi: 10.1016/s0165-0327(02)00008-3

Keysar, B., Lin, S., & Barr, D. J. (2003). Limits on theory of mind use in adults. *Cognition, 89*(1), 25-41. doi: 10.1016/s0010-0277(03)00064-7

Konishi, S., Nakajima, K., Uchida, I., Kikyo, H., Kameyama, M., & Miyashita, Y. (1999). Common inhibitory mechanism in human inferior prefrontal cortex revealed by event-related functional MRI. *Brain, 122*, 981-991. doi: 10.1093/brain/122.5.981

Koster-Hale, J., & Saxe, R. (2013). Theory of Mind: A Neural Prediction Problem. *Neuron, 79*(5), 836-848.

Krause, L., Enticott, P. G., Zangen, A., & Fitzgerald, P. B. (2012). The role of medial prefrontal cortex in theory of mind: A deep rTMS study. *Behavioural Brain Research, 228*(1), 87-90. doi: 10.1016/j.bbr.2011.11.037

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. (2010). Presentation and validation of the Radboud Faces Database. *Cognition & Emotion, 24*(8), 1377-1388. doi: 10.1080/02699930903485076

Leslie, A. M. (1987). Pretense and representation - the origins of theory of mind. *Psychological Review, 94*(4), 412-426. doi: 10.1037/0033-295x.94.4.412

Leslie, A. M., German, T. P., & Polizzi, P. (2005). Belief-desire reasoning as a process of selection. *Cognitive Psychology, 50*(1), 45-85. doi: 10.1016/j.cogpsych.2004.06.002

Leslie, A. M., & Polizzi, P. (1998). Inhibitory processing in the false belief task: Two conjectures. *Developmental Science, 1*(2), 247-253. doi: 10.1111/1467-7687.00038

Lieberman, M. D. (2007). Social Cognitive Neuroscience: A Review of Core Processes. *Annual Review of Psychology, 58*(1), 259-289. doi: 10.1146/annurev.psych.58.110405.085654

Lieberman, M. D., & Cunningham, W. A. (2009). Type I and Type II error concerns in fMRI research: re-balancing the scale. *Social Cognitive and Affective Neuroscience, 4*(4), 423-428. doi: 10.1093/scan/nsp052

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. doi: http://dx.doi.org/10.1016/j.neuroimage.2011.02.067

Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annual Review of Psychology, 62*, 103-134. doi: 10.1146/annurev-psych-120709-145406

Mars, R. B., Sallet, J., Schuffelgen, U., Jbabdi, S., Toni, I., & Rushworth, M. F. (2012). Connectivity-based subdivisions of the human right "temporoparietal junction area": evidence for different areas participating in different cortical networks. *Cerebral Cortex, 22*(8), 1894-1903. doi: 10.1093/cercor/bhr268

McAllister, C. J., Ronnqvist, K. C., Stanford, I. M., Woodhall, G. L., Furlong, P. L., & Hall, S. D. (2013). Oscillatory Beta Activity Mediates Neuroplastic Effects of Motor Cortex Stimulation in Humans. *Journal of Neuroscience, 33*(18), 7919-7927. doi: 10.1523/jneurosci.5624-12.2013

McCleery, J. P., Surtees, A. D., Graham, K. A., Richards, J. E., & Apperly, I. A. (2011). The neural and cognitive time course of theory of mind. *The Journal of Neuroscience, 31*(36), 12849-12854. doi: 10.1523/jneurosci.1392-11.2011

Menzies, T. (1996). Applications of abduction: Knowledge-level modelling. *International Journal of Human-Computer Studies, 45*(3), 305-335. doi: 10.1006/ijhc.1996.0054

Mitchell, J. P. (2007). Activity in Right Temporo-Parietal Junction is Not Selective for Theory-of-Mind. *Cerebral Cortex, 18*(2), 262-271. doi: 10.1093/cercor/bhm051

Mitchell, J. P., & Lacohee, H. (1991). Children's early understanding of false belief. *Cognition, 39*(2), 107-127.

Morris, H. C. (1992). Logical creativity. *Theory & Psychology, 2*(1), 89-107. doi: 10.1177/0959354392021005

Muller, U., Zelazo, P. D., & Imrisek, S. (2005). Executive function and children's understanding of false belief: how specific is the relation? *Cognitive Development, 20*(2), 173-189. doi: 10.1016/j.cogdev.2004.12.004

Nixon, P., Lazarova, J., Hodinott-Hill, I., Gough, P., & Passingham, R. (2004). The inferior frontal gyrus and phonological processing: An investigation using rTMS. *Journal of Cognitive Neuroscience, 16*(2), 289-300. doi: 10.1162/089892904322984571

O'Reilly, R. C. (2010). The What and How of prefrontal cortical organization. *Trends in Neurosciences, 33*(8), 355-361. doi: 10.1016/j.tins.2010.05.002

Olson, I. R., Plotzker, A., & Ezzyat, Y. (2007). The Enigmatic temporal pole: a review of findings on social and emotional processing. *Brain, 130*(Pt 7), 1718-1731. doi: 10.1093/brain/awm052

Pagnucco, M. (1996). *The Role of Abductive Reasoning within the Process of Belief Revision.* Unpublished doctoral dissertation, University of Sydney, Australia.

Pascual-Leone, A., Walsh, V., & Rothwell, J. (2000). Transcranial magnetic stimulation in cognitive neuroscience--virtual lesion, chronometry, and functional connectivity. *Current Opinion in Neurobiology, 10*(2), 232-237.

Perner, J., Aichhorn, M., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Thinking of mental and other representations: The roles of left and right temporo-parietal junction. *Social Neuroscience, 1*(3-4), 245-258. doi: 10.1080/17470910600989896

Perner, J., & Lang, B. (1999). Development of theory of mind and executive control. *Trends in Cognitive Sciences, 3*(9), 337-344.

Perner, J., & 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 Quarterly Journal of Experimental Psychology, 61*(1), 76-89. doi: 10.1080/17470210701508756

Pobric, G., & Hamilton, A. F. D. (2006). Action understanding requires the left inferior frontal cortex. *Current Biology, 16*(5), 524-529. doi: 10.1016/j.cub.2006.01.033

Poldrack, R. A., & Mumford, J. A. (2009). Independence in ROI analysis: where is the voodoo? *Social Cognitive and Affective Neuroscience, 4*(2), 208-213. doi: 10.1093/scan/nsp011

Prado, J., Chadha, A., & Booth, J. R. (2011). The Brain Network for Deductive Reasoning: A Quantitative Meta-analysis of 28 Neuroimaging Studies. *Journal of Cognitive Neuroscience, 23*(11), 3483-3497.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind. *Behavioral and Brain Sciences, 1*(4), 515-526.

Preston, C., & Newport, R. (2008). Misattribution of movement agency following right parietal TMS. *Social Cognitive and Affective Neuroscience, 3*(1), 26-32. doi: 10.1093/scan/nsm036

Quadflieg, S., Turk, D. J., Waiter, G. D., Mitchell, J. P., Jenkins, A. C., & Macrae, C. N. (2009). Exploring the Neural Correlates of Social Stereotyping. *Journal of Cognitive Neuroscience, 21*(8), 1560-1570. doi: 10.1162/jocn.2009.21091

Rakoczy, H., Warneken, F., & Tomasello, M. (2007). "This way!", "No! That way!" - 3-year olds know that two people can have mutually incompatible desires. *Cognitive Development, 22*(1), 47-68. doi: 10.1016/j.cogdev.2006.08.002

Ramnani, N., & Owen, A. M. (2004). Anterior prefrontal cortex: insights into function from anatomy and neuroimaging. *Nature Reviews Neuroscience, 5*(3), 184-194. doi: 10.1038/nrn1343

Ramsey, R., & Hamilton, A. F. D. (2012). How does your own knowledge influence the perception of another person's action in the human brain? *Social Cognitive and Affective Neuroscience, 7*(2), 242-251. doi: 10.1093/scan/nsq102

Ramsey, R., Hansen, P., Apperly, I., & Samson, D. (2013). Seeing it my way or your way: frontoparietal brain areas sustain viewpoint-independent perspective selection processes. *Journal of Cognitive Neuroscience, 25*(5), 670-684. doi: 10.1162/jocn_a_00345

Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: evidence from 14- and 18-month-olds. *Developmental Psychology, 33*(1), 12-21.

Ridderinkhof, K. R., Ullsperger, M., Crone, E. A., & Nieuwenhuiss, S. (2004). The role of the medial frontal cortex in cognitive control. *Science, 306*(5695), 443-447. doi: 10.1126/science.1100301

Ridding, M. C., & Ziemann, U. (2010). Determinants of the induction of cortical plasticity by non-invasive brain stimulation in healthy subjects. *The Journal of Physiology, 588*(13), 2291-2304. doi: 10.1113/jphysiol.2010.190314

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience, 27*, 169-192. doi: 10.1146/annurev.neuro.27.070203.144230

Rose, J. P., Jasper, J. D., & Corser, R. (2012). Interhemispheric interaction and egocentrism: The role of handedness in social comparative judgement. *British Journal of Social Psychology, 51*(1), 111-129. doi: 10.1111/j.2044-8309.2010.02007.x

Rothmayr, C., Sodian, B., Hajak, G., Döhnel, K., Meinhardt, J., & Sommer, M. (2011). Common and distinct neural networks for false-belief reasoning and inhibitory control. *NeuroImage, 56*(3), 1705-1713. doi: 10.1016/j.neuroimage.2010.12.052

Rowe, A. D., Bullock, P. R., Polkey, C. E., & Morris, R. G. (2001). 'Theory of mind' impairments and their relationship to executive functioning following frontal lobe excisions. *Brain, 124*, 600-616. doi: 10.1093/brain/124.3.600

Roxβnagel, C. (2000). Cognitive load and perspective-taking: applying the automatic-controlled distinction to verbal communication. *European Journal of Social Psychology, 30*(3), 429-445. doi: 10.1002/(SICI)1099-0992(200005/06)30:3<429::AID-EJSP3>3.0.CO;2-V

Ruby, P., & Decety, J. (2003). What you believe versus what you think they believe: a neuroimaging study of conceptual perspective-taking. *The European Journal of Neuroscience, 17*(11), 2475-2480.

Rushworth, M. F. S., Buckley, M. J., Behrens, T. E. J., Walton, M. E., & Bannerman, D. M. (2007). Functional organization of the medial frontal cortex. *Current Opinion in Neurobiology, 17*(2), 220-227. doi: 10.1016/j.conb.2007.03.001

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, 40*(6), 859-868.

Sabbagh, M. A., Moses, L. J., & Shiverick, S. (2006). Executive functioning and preschoolers' understanding of false beliefs, false photographs, and false signs. *Child Development, 77*(4), 1034-1049. doi: 10.1111/j.1467-8624.2006.00917.x

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. *Brain, 128*(Pt 5), 1102-1111. doi: 10.1093/brain/awh464

Saxe, R., & Andrews-Hanna, J. R. (n.d.). Retrieved 10/07, 2012, from http://saxelab.mit.edu/stimuli.php

Saxe, R., Brett, M., & Kanwisher, N. (2006). Divide and conquer: a defense of functional localizers. *NeuroImage, 30*(4), 1088-1096; discussion 1097-1089. doi: 10.1016/j.neuroimage.2005.12.062

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. doi: 10.1016/s1053-8119(03)00230-1

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. *Psychological Science, 17*(8), 692-699. doi: 10.1111/j.1467-9280.2006.01768.x

Saxe, R., Schulz, L. E., & Jiang, Y. V. (2006). Reading minds versus following rules: Dissociating theory of mind and executive control in the brain. *Social Neuroscience, 1*(3-4), 284-298. doi: 10.1080/17470910601000446

Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia, 43*(10), 1391-1399. doi: 10.1016/j.neuropsychologia.2005.02.013

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The Future of Memory: Remembering, Imagining, and the Brain. *Neuron, 76*(4), 677-694. doi: 10.1016/j.neuron.2012.11.001

Schilbach, L., Eickhoff, S. B., Rotarska-Jagiela, A., Fink, G. R., & Vogeley, K. (2008). Minds at rest? Social cognition as the default mode of cognizing and its putative relationship to the "default system" of the brain. *Consciousness and Cognition, 17*(2), 457-467. doi: 10.1016/j.concog.2008.03.013

Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E. N., & Saxe, R. (2009). Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. *PLoS One, 4*(3), e4869. doi: 10.1371/journal.pone.0004869

Semendeferi, K., Lu, A., Schenker, N., & Damasio, H. (2002). Humans and great apes share a large frontal cortex. *Nature Neuroscience, 5*(3), 272-276. doi: 10.1038/nn814

Sommer, M., Döhnel, K., Sodian, B., Meinhardt, J., Thoermer, C., & Hajak, G. (2007). Neural correlates of true and false belief reasoning. *NeuroImage, 35*(3), 1378-1384. doi: 10.1016/j.neuroimage.2007.01.042

Souza, M. J., Donohue, S. E., & Bunge, S. A. (2009). Controlled retrieval and selection of action-relevant knowledge mediated by partially overlapping regions in left ventrolateral prefrontal cortex. *NeuroImage, 46*(1), 299-307. doi: http://dx.doi.org/10.1016/j.neuroimage.2009.01.046

Spreng, R. N., Mar, R. A., & Kim, A. S. N. (2009). The Common Neural Basis of Autobiographical Memory, Prospection, Navigation, Theory of Mind, and the Default Mode: A Quantitative Meta-analysis. *Journal of Cognitive Neuroscience, 21*(3), 489-510. doi: 10.1162/jocn.2008.21029

Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience, 10*(5), 640-656. doi: 10.1162/089892998562942

Stuss, D. T., & Alexander, M. P. (2007). Is there a dysexecutive syndrome? *Philosophical Transactions of the Royal Society B-Biological Sciences, 362*(1481), 901-915. doi: 10.1098/rstb.2007.2096

Stuss, D. T., & Benson, D. F. (1984). Neuropsychological studies of the frontal lobes. *Psychological Bulletin, 95*(1), 3-28. doi: 10.1037//0033-2909.95.1.3

Stuss, D. T., Gallup, G. G., & Alexander, M. P. (2001). The frontal lobes are necessary for 'theory of mind'. *Brain, 124*, 279-286. doi: 10.1093/brain/124.2.279

Surtees, A. D. R., & Apperly, I. A. (2012). Egocentrism and Automatic Perspective Taking in Children and Adults. *Child Development, 83*(2), 452-460. doi: 10.1111/j.1467-8624.2011.01730.x

Thomas, R. C., & Jacoby, L. L. (2013). Diminishing adult egocentrism when estimating what others know. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(2), 473-486. doi: 10.1037/a0028883

Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences, 12*(12), 455-460. doi: 10.1016/j.tics.2008.10.001

Uddin, L. Q., Molnar-Szakacs, I., Zaidel, E., & Iacoboni, M. (2006). rTMS to the right inferior parietal lobule disrupts self-other discrimination. *Social Cognitive and Affective Neuroscience, 1*(1), 65-71. doi: 10.1093/scan/nsl003

van der Meer, L., Groenewold, N. A., Nolen, W. A., Pijnenborg, M., & Aleman, A. (2011). Inhibit yourself and understand the other: Neural basis of distinct processes underlying Theory of Mind. *NeuroImage, 56*(4), 2364-2374. doi: 10.1016/j.neuroimage.2011.03.053

Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping, 30*(3), 829-858. doi: 10.1002/hbm.20547

Van Overwalle, F. (2011). A dissociation between social mentalizing and general reasoning. *NeuroImage, 54*(2), 1589-1599. doi: 10.1016/j.neuroimage.2010.09.043

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage, 48*(3), 564-584. doi: 10.1016/j.neuroimage.2009.06.009

Vanderwert, R. E., Fox, N. A., & Ferrari, P. F. (2013). The mirror mechanism and mu rhythm in social development. *Neuroscience Letters, 540*, 15-20. doi: 10.1016/j.neulet.2012.10.006

Verbruggen, F., Aron, A. R., Stevens, M. A., & Chambers, C. D. (2010). Theta burst stimulation dissociates attention and action updating in human inferior frontal cortex. *Proceedings of the National Academy of Sciences, 107*(31), 13966-13971. doi: 10.1073/pnas.1001957107

Vernet, M., Bashir, S., Yoo, W. K., Oberman, L., Mizrahi, I., Ifert-Miller, F., . . . Pascual-Leone, A. (2013). Reproducibility of the effects of theta burst stimulation on motor cortical plasticity in healthy participants. *Clinical Neurophysiology*. doi: 10.1016/j.clinph.2013.07.004

Verschuere, B., Schuhmann, T., & Sack, A. T. (2012). Does the inferior frontal sulcus play a functional role in deception? A neuronavigated theta-burst transcranial magnetic stimulation study. *Frontiers in Human Neuroscience, 6*, 284. doi: 10.3389/fnhum.2012.00284

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., . . . Zilles, K. (2001). Mind Reading: Neural Mechanisms of Theory of Mind and Self-Perspective. *NeuroImage, 14*(1), 170-181. doi: 10.1006/nimg.2001.0789

Watkins, K., & Paus, T. (2004). Modulation of motor excitability during speech perception: The role of Broca's area. *Journal of Cognitive Neuroscience, 16*(6), 978-987. doi: 10.1162/0898929041502616

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

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

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences of the United States of America, 107*(15), 6753-6758. doi: 10.1073/pnas.0914826107

Young, L., Dodell-Feder, D., & Saxe, R. (2010). What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. *Neuropsychologia, 48*(9), 2658-2664. doi: 10.1016/j.neuropsychologia.2010.05.012

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

APPENDIX 1

Participant Recruitment and Selection

Online Screening

Due to the nature of the experimental tasks and the equipment used in the experiments presented in this thesis, strict exclusionary criteria were applied. This was so that the safety of the participant group could be ensured, and to identify a maximally homogeneous (i.e. considered behaviourally and neurally typical) participant group.

The opportunity to participate in each experiment was advertised through the University of Birmingham's online Research Participation Scheme. Before registering their interest to take part, prospective participants were asked to read the below 10 questions. If they answered yes to any of the questions, they were asked not to continue with their online application to participate.

1. Are you left handed?
2. Is your native language anything other than English?
3. Are you bilingual?
4. Are you dyslexic?
5. Are you pregnant?
6. Are you claustrophobic?
7. Do you have any metal or surgical implants (e.g. pacemakers, surgical clips etc)?
8. Do you suffer from any neurological or psychiatric condition?

9. Have you ever suffered a stroke, or any other brain injury?
10. Have you been diagnosed with an Autistic Spectrum Disorder or ADHD?

Questions 1 to 3 screen for issues which may affect the homogeneity of the neural data collected. There are data which suggest different behavioural profiles and functional brain organisation in individuals who are left handed or bilingual compared with typical controls. Questions 2 and 4 reflect the requirement to ensure reading proficiency in line with the task requirements; the localizer task, for example, which was used alongside the majority of the paradigms presented in this thesis, requires confidence with written English. Questions 5 to 7 screen for safety/comfort contraindications which may preclude the individual from taking part in an Magnetic Resonance Imaging (MRI) experiment. Note that for the experiment outlined in Chapter 4, additional screening questions were included (see Appendix 2), which reflect the School of Psychology's standard ethical protocol for studies involving Transcranial Magnetic Stimulation (TMS). Questions 8 to 10 screen for conditions which may affect the homogeneity of the data, in terms of reflecting the behavioural profile and functional/cortical organisation of typical participants.

Behavioural Pre-screen

After registering their interest through the online Research Participation Scheme, all participants physically attended a pre-screen. All were asked to reconfirm their status regarding the previously outlined 10 questions. The main purpose of the pre-screen, however, was to identify participants who were able to perform the task of interest; any individual who performed below chance was not invited to participate in the neuroimaging/TMS experiments.

This was on the premise that it was only possible to negotiate the tasks by engaging a Theory of Mind (ToM), whereas incorrect trials reflect heterogeneous causes. In order to take part in the experiments outlined in Chapters 2 to 4, participants needed to be able to perform the main experimental task, including any filler/catch trials, to above chance at $p < 0.05$ (calculated for a binomial distribution in all cases). Before attempting any trials, participants were taken through an interactive training programme which outlined the task and included practise trials with feedback on their performance. No behavioural pre-test was required for the experiment outlined in Chapter 5. During the pre-test for all experiments outlined in this thesis, participants completed a handedness measure (Annett, 1970) and a simple reading scale (WRAT 3) which required them to read a selection of words aloud. The latter measure was used as a coarse method to identify and exclude individuals who may have a reading disability from participating. This was particularly important as the localizer task, which was regularly used alongside the main task of interest, was not part of the behavioural pre-test.

APPENDIX 2

Additional Screening for TMS

To ensure participant safety, the University of Birmingham's School of Psychology requires that the below information is collected from participants prior to them participating in any TMS study. A 'yes' response indicates a possible TMS contraindication, which would be reviewed on a case-by-case basis.

Have you ever suffered from any neurological or psychiatric conditions? . . . . . YES / NO
If YES please give details (nature of condition, duration, current medication, etc)

. . . . . .
Have you ever suffered from epilepsy, febrile convulsions in infancy
or had recurrent fainting spells? . . . . . YES / NO
Does anyone in your immediate or distant family suffer from epilepsy? . . . . . YES / NO
If YES please state your relationship to the affected family member.

. . . . . .
Do you suffer from migraine? . . . . . YES / NO
Have you ever undergone a neurosurgical procedure (including eye surgery)? . . . . . YES / NO
If YES please give details.
Do you currently have any of the following fitted to your body? . . . . . YES / NO
- Heart pacemaker
- Cochlear implant
- Medication pump
- Surgical clips

Are you currently taking any unprescribed or prescribed medication? . . . . . YES / NO
If YES please give details.
. . . . . .
Are you currently undergoing anti - malarial treatment? . . . . . YES / NO
Have you drunk more than 3 units of alcohol in the last 24 hours? . . . . . YES / NO
Have you drunk alcohol already today? . . . . . YES / NO
Have you had more than one cup of coffee, or other sources of caffeine,
in the last hour? . . . . . YES / NO
Have you used recreational drugs in the last 24 hours? . . . . . YES / NO
Did you have very little sleep last night? . . . . . YES / NO
Have you already participated in a TMS experiment today? . . . . . YES / NO
Have you taken part in 2 or more TBS or tDCS experiments in the last 6 months? . . . YES / NO
Are you taking any prescribed drugs (prescribed by your GP or a hospital)? . . . . . YES / NO
Is there any chance that you could be pregnant? . . . . . YES / NO
Are you left or right handed? . . . . . Left / Right
Date of Birth _____/_____/_____