In press in Cognition

Beyond Simulation-Theory and Theory-Theory: Why social cognitive neuroscience should use its own concepts to study “Theory of Mind”

Ian A. Apperly

University of Birmingham, UK

Correspondence concerning this article should be addressed to Ian Apperly, School of Psychology, University of Birmingham, Edgbaston, Birmingham, B15 2TT, UK. E-mail: i.a.apperly@bham.ac.uk

Abstract

The debate between Simulation-Theory (ST) and Theory-Theory (TT) provides the dominant theoretical framework for research on “theory of mind” (ToM). Behavioural research has failed to provide clear methods for discriminating between these theories, but a number of recent studies have claimed that neuroimaging methods do allow key predictions of ST and TT to be tested. In the current paper it is argued that neuroimaging studies have not in fact provided any data that discriminates between ST and TT accounts of propositional attitude ascription, and moreover that it is uncertain that they will in the future. However, it is also argued that the fault lies with the ST/TT debate, not with the methods and concepts of neuroimaging research. Neuroimaging can certainly contribute to our understanding of ToM, and should contribute to the project of developing theoretical models more firmly grounded in specific cognitive and neural processes than ST or TT.

Social cognitive neuroscience gives us exciting new ways to study how humans and other animals explain and predict behaviour in terms of mental states. This emerging discipline can usefully inherit many concepts and paradigms from twenty five years of developmental, comparative and theoretical work into theory of mind (see e.g., Apperly, Samson & Humphreys, 2005; Frith & Frith 2003; Saxe, 2006). However, I will argue that an important exception is the longstanding debate between "Theory-Theory" (TT) and "Simulation-Theory" (ST) accounts of Theory of Mind (ToM). Despite claims to the contrary, social cognitive neuroscience has been no more successful than behavioural approaches in producing clear evidence to discriminate between these theories. Although better evidence may be forthcoming, it may also be that the debate between TT and ST is not the most useful theoretical framework for generating predictions or interpreting data in investigations of ToM. I will suggest that social cognitive neuroscience is already equipped with the right conceptual tools for generating empirically tractable hypotheses that will be a more reliable way of advancing our understanding of ToM.

Different ways of explaining behaviour.

"Theory of mind" or folk psychology is the ability to treat agents as the owners of unobservable mental states - beliefs, desires and the like - and to explain and predict the behaviour of agents in terms of such mental states[1]. To illustrate what is distinctive about a ToM appraisal of an agent's behaviour, consider the following:

1) George likes to go to the gym in the morning, but he forgot it was closed on Mondays, so when he got there he just went straight to work.

2) George usually goes to the gym in the morning but when he got there today it was closed, so he just went straight to work.

These sentences describe the same objective event sequence. They both allow predictions to be made about behaviour in the future. However, sentence 1 gives us additional information about George's internal mental states allowing us to predict that he will be annoyed that he forgot that the gym would be closed and disappointed that he will not be able to exercise. In contrast, although sentence 2 allows this interpretation, it also allows a variety of others. For example, George might only be going to the gym out of obligation, so was relieved that the gym was closed and went to work feeling happy. Thus, a theory of mind appraisal of behaviour can give additional purchase on the problem of understanding the behaviour of agents. This is a principal reason why researchers have tried to understand how behaviour is appraised in terms of ToM concepts, and highlights why it is important to distinguish such cases from the broader category of explanations and predictions made in terms of generalisations over observable features of behaviour.

Theory-Theory and Simulation-Theory.

Philosophers distinguish two very general accounts of theory of mind, which have provided the dominant interpretive frameworks for empirical investigations. Theory-Theory (TT) accounts propose that theory of mind abilities are constituted by a set of concepts (belief, desire etc.) and governing principles about how these concepts interact (e.g., people act to satisfy their desires according to their beliefs). The proposed status of these concepts and principles varies widely, from symbols and processing rules in sub-personal Language of Thought (for a discussion see Stich & Nichols, 1992), to a set of personal-level notions and hypotheses to which we have explicit access (e.g., Gopnik & Wellman, 1992; Gopnik & Melzoff, 1997). What such accounts share, however, is the assumption that these concepts and principles constitute a causal "theory" of how an agent's mental states interact to generate behaviour, and that this theory, in combination with appropriate initial information

about the agent, is the means by which we formulate explanations and predictions about mental states and behaviour (see Figure 1).

Simulation-Theory (ST) accounts were developed as a sceptical response to the claim that TT explains all instances of ToM reasoning (e.g., Gordon, 1986; Heal, 1986). Simulationists note that biology ensures that our own minds will have processes for the fixation of beliefs, the formation of desires and other processes involving mental states that are essentially similar in their causal properties to the same processes in the minds of others. This being the case, at least some of the work involved in thinking about another mind could be achieved by using one's own mind as a model. A necessary precursor to such a process would be to work out the target's set of initial mental states (at least some of which would be different from one's own). The causal interactions of these states could then be modelled by using one's own mind "off-line," effectively as a simulator of the target's mind. The outputs from this process would be de-coupled from their usual role in governing our own behaviour, and would instead form predictions about the mental states and behaviour of the target (see Figure 1). The simulationists' argument is that, given this possibility, it is unnecessary and unparsimonious to suppose that our ability to explain and predict behaviour requires an exhaustive *theory* of the causal interactions of mental states. Like TT, simulation accounts vary widely in form, from simulation as a process of deliberate, personal-level introspection and projection (e.g., Goldman, 1989; Harris, 1989), to automatic sub-personal "resonance" between the simulating agent and the target (e.g., Gallese & Goldman, 1998).

Although ST and TT were originally viewed as mutually exclusive accounts of ToM many authors now argue for a hybrid account in which both Simulation and Theory play a role (e.g., Carruthers & Smith, 1995; Currie & Ravenscroft, 2002; Nichols & Stich, 2003). Part of the motivation for this is that both ST and TT seem to have compelling cases in their favour. It is widely agreed that when we anticipate someone else's judgement about the grammaticality

of a sentence we use our own (non-theoretical) grammatical intuitions, and that this provides a compelling case of simulation (e.g., Harris, 1989). On the other hand, cases where people make systematic errors in their predictions about the decisions of others (and, indeed, predictions about their own decisions) are seen by many as good evidence that at least some ToM judgements are informed by a theory about how the mind works, and that this theory is sometimes wrong (e.g., Saxe, 2005; Nichols & Stich, 2003). Hybrid ST/TT accounts are an important theoretical shift in the ST/TT debate. However, the challenge for experimental investigators remains essentially similar: for any given judgement about another person, can experiments be devised that provide clear evidence for a role for either simulation or theoretical reasoning?

In many cases, behavioural evidence has proved inconclusive at discriminating ST from TT, because although data may allow some versions of ST or TT to be excluded, other versions of either theory are able to explain the findings (e.g., Carruthers & Smith, 1996; Saxe, 2005; Stich & Nichols, 1997). Theorists despairing at the possibility of discriminating ST and TT with behavioural data have sometimes hinted that neuroimaging has the potential to provide clearer evidence (e.g., Stich & Nichols, 1997). In recent years there have been many neuroimaging investigations of ToM (for recent reviews see e.g., Frith & Frith, 2003, 2006). However, although ST and TT often form part of the interpretive background in these studies, only a few investigations have taken up Stich and Nichols' (1997) challenge to use neuroimaging in a direct test of predictions arising from ST and TT for ToM (that is to say, for propositional mental states, rather than emotions or actions). In the current paper I focus on four papers that make the strongest claims about testing these predictions. I will examine two strategies that have been used: Comparing the neural activation for predicting judgements or actions about self versus other; and examining the degree to which activations when making judgments about others are modulated by perceived similarity to self. I will argue that both

approaches founder and that understanding why reveals confusion about critical concepts such as “self”, self-other similarity and, indeed, what counts as a “mental state” in work on ToM. Careful attention to these conceptual issues will surely enhance the contribution that neuroscientific approaches will make to our understanding of ToM, and just may allow neuroscience to shed light on the ST/TT debate.

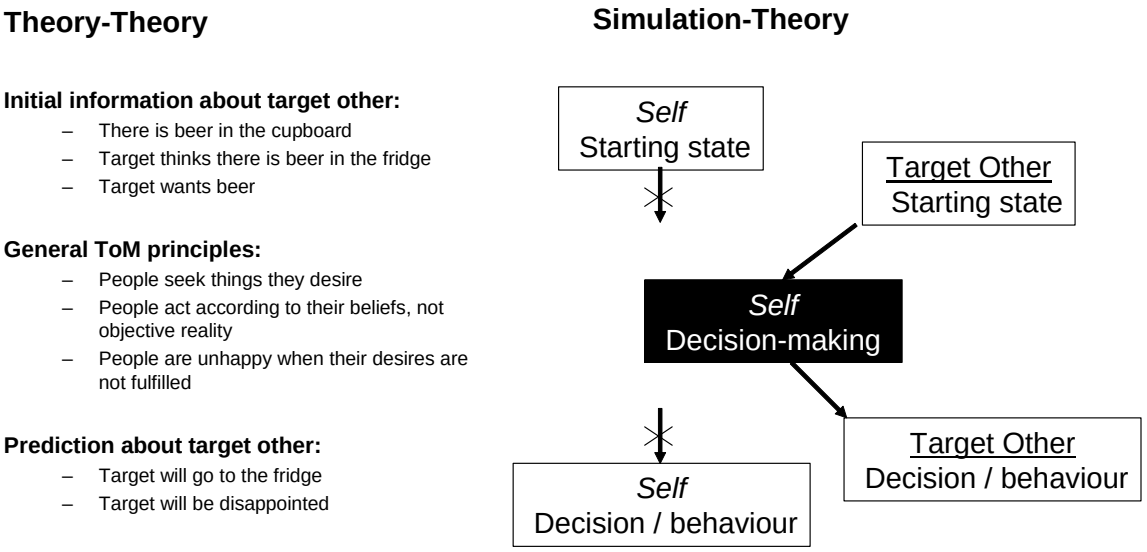

Figure 1. *Schematic representation of Simulation-Theory and Theory-Theory accounts of Theory of Mind.*
Theory-Theory: Starting with initial information about the target's beliefs and desires the agent uses general ToM principles to generate a prediction about the target's future mental states and behaviour. Simulation-Theory: The agent first takes their own decision-making system off-line from its usual role in guiding the agent's behaviour. Starting with initial information about the target's beliefs and desires, this information is fed into the agent's own decision-making system, which generates a decision or behavioural output that can be taken as a prediction of the decision or behaviour of the target other. Whereas Theory-Theory requires that the General ToM principles constitute an exhaustive account of the causal workings of the mind, Simulation-Theory holds that at least some of these principles can remain implicit in the processes for Self decision-making. Both theories require the agent to have appropriate initial information about the target. In either case specifying this initial information may depend upon sophisticated reasoning and heuristics such as perceived similarity to self as well as more automatic processes of social categorisation and person perception.

Comparing neural activation for judgements about self and other.

According to ST, predictions about what a target person will think or do depend, at least in part, upon using our own mind to simulate the target's mental processes. In contrast, TT suggests that predictions about others depend upon a different set of processes (involving concepts and general principles) from those involved when our own beliefs, desires and intentions form the basis for our own behaviour. It follows that an appropriate comparison of neural activation for "self" and "other" processes could provide strong evidence to discriminate ST from TT. If common activation was observed for self and other processes this would seem to favour ST over TT, whereas the absence of common activation would favour TT over ST. Several studies have pursued this strategy. However, although they each produce informative results, I will suggest that each one fails to discriminate between ST and TT for ToM, either because they make the wrong kind of self-other comparison or because they fail to make this comparison for ToM content (i.e., mental states such as beliefs and desires).

Ramnani and Miall (2004) conducted a study that has the right form of self-other comparison but lacks the necessary ToM content. Participants were instructed to press buttons according to a simple set of rules. On each trial the participant was shown a coloured geometric shape. The particular colour of the shape determined whether a response should be given by the participant, a second human player or the computer. The particular shape determined which button response should be given by whoever was due to respond on that trial. Feedback on the participant's monitor allowed the participant to see which button had in fact been pressed on each trial and as well as responding themselves, participants were asked to monitor for errors made by the other human player or the computer. This manipulation was designed to encourage participants to anticipate the actions of the second human player and the computer, making it possible to compare the neural activation for anticipating the other person's actions versus the participant preparing to respond

themselves. Neural activation was observed in premotor cortex for both “self” and human “other” trials, consistent with the idea that participants were using their own motor systems to anticipate the other person’s actions. Critically however, these regions of activation did not overlap, suggesting that distinct neural systems were being used to anticipate the actions of self and other. The authors concluded that this favours TT over ST.

This study clearly has the right *form* to pit ST against TT insofar as it compares the participant themselves making a decision (using their own 1st person mental machinery) with the participant anticipating this same decision in someone else. The critical problem is that the task does not appear to have any ToM *content*. Recall from earlier that there are different ways of explaining and predicting behaviour, and that the distinctive feature of ToM explanations and predictions is that they depend upon reasoning about an agent’s internal mental states. The task employed by Ramnani and Miall (2004) requires participants to anticipate responses for self and other according to a simple set of stimulus-response mappings. It is not necessary to consider beliefs, desires, intentions or other mental states, in either the self condition or the other condition. Thus, although these findings lead to interesting conclusions about the cognitive and neural basis of action prediction, they do not warrant any conclusions about ToM.

Grezes, Frith and Passingham (2004a; see also Grezes, Frith & Passingham, 2004b) conducted studies that came closer to having ToM content, but lack the appropriate form of self-other comparison to test ST against TT. Grezes et al. (2004a) examined participants’ ability to infer the beliefs of an agent performing a simple action. Stimuli were created by filming participants as they picked up boxes of different weights. On most occasions participants were correctly forewarned about whether the box was heavy or light, but on a small proportion of trials they were misled to expect the box to be heavy when it was in fact

light, or vice versa. At test participants were shown videos of themselves or other participants, and were asked to judge whether the person in the video had a correct or incorrect expectation about the weight of the box. Behaviourally, participants were above chance at discriminating correct from incorrect actions. Neural activation was observed bilaterally in premotor cortex (plus left frontal operculum, left intraparietal sulcus and right cerebellum) when participants judged videos of themselves *and* when they judged videos of other people, though the responses in these areas were significantly faster for self judgements than for other judgements. The authors argue that this common activation for self and other judgements is evidence in favour of ST because participants are using the same premotor circuits for modelling actions for self and other. Activations were also observed in superior temporal sulcus, dorsomedial frontal cortex or paracingulate cortex, orbital frontal cortex and nearby anterior insula and cerebellum for three contrasts: between trials where the agent actually had a true versus a false expectation (regardless of the participants' judgement); between trials where the participant *judged* the agent to have a true versus a false expectation (regardless of the agent's actual expectation); and between trials where the participant made correct versus incorrect judgements about an agent who had a false expectation. The authors argue that this pattern reflects processing of the discrepancy between the participants' expectations about the movements of the agent (based on forward modelling in their premotor circuit) and the movements that participants actually observe, on the basis of which participants inferred whether the agent had a correct or incorrect expectation about the weight of the box.

There are two reasons why this study does not discriminate between ST and TT. First, although participants were explicitly asked to judge whether the agent had a true or false expectation about the weight of the box, there is no evidence that related theory of mind content played any role in the observed patterns of neural activation. As for Ramnani and

Miall (2004) the neural activation in premotor circuits for self and other judgements reflected processing of actions, not processing of mental states such as beliefs, desires or intentions. The other key contrast, between neural activation for true versus false belief trials, or between correct and incorrect judgements about false belief trials, should effectively subtract out any activation associated with belief reasoning per se. The remaining activation may, as the authors suggest, be associated with detection of a discrepancy between expected and observed behaviour, which is an interesting conclusion in its own right. But the absence of ToM content in this contrast means that it could not tell us about ST or TT.

Second, unlike Ramnani and Miall (2004), this study does not make the right form of comparison between activation for "self" and "other" to test hypotheses about ST versus TT. ST holds that the cognitive processes that we use to perform a task for ourselves can serve as a useful model for the cognitive processes that other people would use to solve the same task. The necessary "self" condition in the Grezes et al. (2004) study would have had to identify of the neural activation associated with the participant actually having a current, 1st person true or a false expectation. Instead the "self" condition in this study involved recording activation when participants made 3rd person judgements about themselves at an earlier point in time. This activation was then compared with that observed in an "other" condition in which participants made similar judgements about another agent performing the same lifting task. This yields an interesting finding, suggesting that participants are quicker to model observed actions when they are the observed agent than when the agent is another person. This processing advantage for perceiving one's own actions may contribute importantly to our ability to distinguish between our own actions and those of others. However, no conclusions about ST or TT follow from this finding.

A study by Vogeley, Bussfield, Newen et al. (2001) does have ToM content in its critical comparisons, but these comparisons lack the necessary form to speak to the ST / TT debate. These authors used fMRI to measure neural activation while participants read short stories designed to encourage the adoption of either self- or other-perspective while making either physical or social ("theory of mind") judgements. Theory of mind stories required a judgement about the behaviour of a character where the participant had to take into account the character's beliefs, desires or intentions, whereas physical stories required a judgement about physical causality, such as how an animal running across a room might cause the activation of a burglar alarm (see e.g., Fletcher, Happe, Frith et al., 1995). For other-perspective judgements all of the story characters were fictional (e.g., "A burglar who has just robbed a shop is making his getaway. As he is running home, a policeman on his beat sees him drop his glove…."). When participants were asked to make theory of mind judgements, these judgements were about the fictional characters such as the burglar and the policeman. For self-perspective judgements the stories were modified so that the participant themselves featured in the narrative, and was referred to by 1st person pronouns (e.g., "…A burglar who has just robbed a shop is making his getaway. He has robbed your store. But you cannot stop him."). "Self-perspective" theory of mind judgements were about what the participants themselves would do, say or think in these fictional scenarios.

Regions of right anterior cingulate cortex and left temporopolar cortex were more activated for theory of mind judgements than for physical causality judgements, irrespective of whether these judgements were made for self or other. However, a region of right prefrontal cortex was more activated for the self-perspective theory of mind condition than for any other condition. The authors conclude that the existence of shared neural activation for self- and other-perspective theory of mind judgements favours ST, whereas evidence of distinct processes favours TT, and argue for a hybrid ST/TT account.

Unlike the studies reviewed above, Vogeley et al.'s (2001) comparisons undoubtedly have ToM content because participants are required to make judgements about story characters' beliefs, desires or intentions, and, unlike Grezes et al., (2004), activation in these conditions was contrasted with conditions that lacked ToM content, so the effects of this content were not subtracted out. However, in an analogous way to Grezes et al. (2004), the study lacks the necessary form of self-other comparison to pit ST against TT. Vogeley et al.'s "self" condition involved participants making 3rd person ascriptions of mental states to a hypothetical self in a fictional scenario, and this was contrasted with similar 3rd person ascriptions to a fictional "other" in similar scenarios. This generates an interesting comparison that identifies both common and distinct neural activation involved in 3rd person mental state ascription for self and other. The possible constraints that this places on functional accounts of ToM will be discussed later. The critical point for now, however, is that this comparison cannot test ST's claim that participants use the causal processes of their own 1st person cognition to model the same processes in another mind – to do so would have required a "self" condition in which participants *had for themselves* beliefs, desires or intentions that were relevantly similar to those ascribed to others in a 3rd person condition. Third person ascription of mental states to one's hypothetical self is clearly not the same as being in those mental states one's self. Thus, no conclusions about ST versus TT follow from this study.

The above studies illustrate conceptual issues that need to be addressed if neuroimaging methods are to be used to test between ST and TT accounts of ToM.

Concepts of "self". One problem is that many processes that clearly do involve or relate to the "self" are confusingly diverse, and may be the wrong kind of process for pitting ST against TT. The "self" in Ramnani & Miall (2004) was the experimental participant in the 1st person in real-time, using the rules of the game to prepare their own response to the stimulus

on a given trial. The "self" in Grezes et al. (2004) was the experimental participant, but in the past and perceived from a 3rd person perspective. The "self" in Vogeley et al. (2001) was constructed and projected into a hypothetical scenario and perceived from the 3rd person perspective of a reader of the scenario. (And in the studies reviewed below (Mitchell, Banajii & Macrae, 2005; Mitchell, Macrae & Banaji, 2006) we encounter yet another "self": the participant's own self-concept perceived from a 3rd person introspective perspective.) Clearly, these studies are all working with legitimate concepts of "self", but these concepts are not equivalent. It would be truly surprising if such diverse studies produced convergent evidence about the cognitive and neural basis of "self" and self-related processes, let alone the role of "self" processes in predicting or explaining "others". More importantly for the current discussion, just because each study has a condition corresponding to a legitimate concept of "self", it does not follow that we are dealing with the right kinds of "self" processes to test ST against TT. On this point, the solution seems clear – at least in principle. The right kinds of "self" processes are those where participants are *having for themselves* (in the 1st person, in real time) the same kinds of causally interacting beliefs, desires and intentions that are presumed to be present in the "other" person whose behaviour the participant must predict or explain in the comparable "other" condition. Ramnani and Miall's (2004) study is the only one of the above that employs such a comparison. As discussed below, the problem with this study (for testing ST against TT) is that it does not have appropriate ToM content.

ToM content. As alternative accounts of ToM, ST and TT are concerned with explaining how participants explain and predict the behaviour of others on the basis of causally interacting mental states (such as beliefs, desires and intentions). Thus, for neuroimaging evidence to distinguish between ST and TT accounts of ToM, the task in which

participants are engaged must involve this kind of information processing, and do so for both self and other. The studies reviewed above raise two important problems about ToM content.

The first problem is that varying criteria for what counts as a "ToM task" mean that it is difficult to be confident about what processes are being imaged in different studies. Vogeley et al.'s (2001) study clearly did involve participants making specific inferences about the beliefs of characters in the stories. In contrast, Ramnani & Miall's (2004) task only involved participants behaving or predicting behaviour on the basis of conditional rules, so may not have involved processing beliefs, desires or intentions at all. Grezes et al.'s (2004) study required participants to judge a target's (true or false) expectation about the weight of a box they were lifting. However, although the target's expectation logically follows from their belief, it is not certain that participants actually made these belief ascriptions because the target's expectations could be judged correctly simply by interpreting their motor behaviour. (As we shall see below, Mitchell et al.'s (2005, 2006) studies required participants to gauge trait-related attributes, such as likes, dislikes and opinions of others, but not trait-independent states of belief or knowledge.) When compared with each other it is clear that these studies adopt very different criteria for what counts as a "theory of mind" task. It seems likely that such diverse tasks might depend upon importantly different functional and neural processes. Understanding these processes in future work will depend upon developing much more explicit and detailed task analyses for different kinds of ToM task.

The second problem about ToM content is related, but deeper, and much more difficult to solve. It was concluded above that for neuroimaging evidence to distinguish between ST and TT for a given ToM process a study must include a "self" condition in which participants *have for themselves* (in the 1st person, in real time) the relevant mental states. However, this leads to a serious problem because of practical difficulty and deep theoretical controversy about when and whether we can consider someone to have a belief, a desire or an intention

(e.g., Baker, 1995; Bermudez, 2005; Churchland, 1986; Dennett, 1987; Fodor, 1975; Searle, 1983).

Consider, for illustration, the Grezes et al. (2004) study, where participants made 3rd person judgements about an actor's beliefs about the weight of a box they were lifting. The necessary "self" condition would require the participants to be *having for themselves* the mental states that they would be ascribing to the other person in the "other" condition. Setting practical issues aside for a moment, it is tempting to think that this could be achieved if neural activation were monitored while the participants themselves were told the supposed weight of the box and then found out the truth or falsity of this belief when they picked it up. However, to interpret such neural activation with confidence we would have to be certain that under these conditions the participant actually *had* a belief about the weight of the box, and here there is substantial room for doubt.

Of course, in one sense it seems perfectly clear that the participant's motor system has an implicit belief about the weight of the box when the motor plan for picking up the box is formulated. However, it is also clear that the status of these beliefs is complicated. For example, the ability to perform accurate, visually guided, object-directed actions (based upon "motor beliefs") can doubly dissociate from the ability to make accurate judgements about the size, shape or orientation of objects on the basis of the same incoming visual information (e.g., Milner & Goodale, 1992). That is to say, "motor beliefs" are very different from the "personal-level" beliefs we routinely ascribe to people in our everyday folk psychology, and in fact a single participant can hold contradictory motor beliefs and personal-level beliefs. As it happens, it is personal-level beliefs that have been the primary subject for the ST/TT debate. But this does not matter so much as the fact that deciding whether or not a participant *has* a belief is no simple matter even in this simple case.

Matters are no easier if we restrict ourselves only to considering personal-level beliefs. For example, just because the participant is told the weight of the box by the experimenter this does not mean that they "believe" what they are told. They may judge that the experimenter has misinformed them and so believe the opposite of what they are told. Alternatively, they may not actively disbelieve the experimenter, but may not be committed to the truth of what they have been told. Indeed, even if they do "believe" the experimenter, they may not formulate an explicit belief about the weight of the box. Finally, even when the participant picks up the box it is unclear whether this actually leads the participant to formulate a particular belief about the box's weight or just to have the dispositional tendency (perhaps as a motor belief) for such a belief to be formulated.

In sum, a test of ST's apparently simple prediction of overlapping neural processes for "self" and "other" ToM processes depends upon a scientific account of where, how and when mental states such as beliefs and desires are formulated in the 1st person case. This is an unusually high demand. Many theories in psychology assume that people have mental states such as beliefs, for example: cognitive dissonance in social psychology (Festinger, 1957); the "curse of knowledge" in decision-making (Camerer, Lowenstein, & Weber, 1989); core knowledge in infancy (e.g., Spelke & Kinzler, 2007). However, the use that these theories make of such mental states does not typically depend upon resolving any uncertainty about whether a person implicitly believes something, whether they are disposed to believe something, or whether an explicitly held belief is in long term memory or is in fact making a current, direct contribution to behaviour. Testing ST's apparently simple prediction of overlapping neural processes for "self" and "other" ToM processes does depend upon resolving such uncertainties because it is necessary to be able to associate a pattern of neural activation unambiguously with a certain belief, desire or intention in order to be able to compare this with activation when the participant ascribes the same belief, desire or intention

to someone else. As a result, testing ST against TT by comparing “self” and “other” neural activation is rather more difficult than might have been expected. Indeed if future work fails to solve the more difficult problem of identifying conditions under which we can be sure that a participant is in a current state of believing, intending or desiring then testing ST against TT in this way will be impossible.

Dependence of judgements about others on perceived similarity to self.

In recent studies it has been suggested that functional neuroimaging can provide data to discriminate ST from TT by showing that the neural activity for third-person ToM judgements is modulated by perceived similarity to self (e.g., Mitchell, Banajii & Macrae, 2005; Mitchell, Macrae & Banaji, 2006; Frith & Frith, 2006; Saxe & Wexler, 2005). This argument has been made most strongly by Mitchell and colleagues in two studies, and it is their argument and data that will be discussed here. There are two components to their contention. First, Mitchell et al. (2005) argue that ST equates with self-reflection and projection: that is to say, the simulator first imagines what they themselves would do in a given set of circumstances and then uses this as a model for what another person would do. Second, they argue that “…simulation accounts of mental state attribution suggest that perceivers only use self-reflection as a strategy to predict the mental states of others when these individuals are in some way similar to self” (p 1307). On this argument, it would count in favour of ST if it could be shown that a neural region involved in self-reflection was recruited for judgements about other people, and modulated by the perceived similarity between self and other.

To test this hypothesis Mitchell, Banajii and Macrae (2005; see also Mitchell, Macrae & Banaji, 2006) measured neural activation while participants made social or non-social judgements about photographs of faces. Social judgements required participants to rate how

pleased the subject of the photograph was to have his or her photograph taken. Non-social judgements required participants to rate the photographs for the symmetricality of the subject's face. In a post-test, participants rated perceived similarity between themselves and the person in each photograph. The critical finding was activation in a region of ventral medial prefrontal cortex (mPFC) – an area activated in independent studies of self-reflection (e.g., Gusnard, Akbudak, Shulman, & Raichle, 2001; Vogeley, May, Ritzl et al., 2004) – that was selective for social rather than non-social judgements, and proportionately more active for social judgements about individuals rated similar rather than dissimilar to self. Mitchell, Macrae and Banaji (2006) replicated this effect, and additionally found the inverse pattern for dorsal mPFC, which was more active for social judgements about dissimilar others.

This is clearly an interesting result, providing evidence for patterns of neural activation that are consistent with the use of introspection to predict the attitudes of other people, in a way that is sensitive to how similar to the self those other people are perceived to be. The conditionalisation of neural activation during social judgements according to perceived similarity between self and other is likely to prove a very useful tool for investigating the neural basis of social cognitive processes. Critically, however, this approach cannot provide data to discriminate ST from TT. The reason for this rests with the two claims on which Mitchell et al's argument is based.

First, the equation between ST and introspection is incorrect. Although some authors have suggested that simulation amounts to a process of introspection and projection (e.g., Goldman, 1989) there is no consensus either that simulation requires introspection (Gallese & Goldman, 1998) or that introspection depends upon simulation rather than reasoning with mental state concepts and rules or some independent mental mechanism (e.g., Nichols & Stich, 2006). Perhaps more importantly, positive evidence for the involvement of introspection in a ToM process is entirely neutral for the ST/TT debate. As is clear from Figure 1, *both* ST

and TT require participants to construct an appropriate set of starting mental states for the person about whom they wish to make a prediction or explanation. Introspection upon what the participant would think or feel themselves is surely a potential source of these starting mental states, but this is equally true for ST and TT, meaning that Mitchell et al.'s evidence for the engagement of a neural system for introspection counts equally well for ST and TT. The *distinctive* feature of ST as opposed to TT is the possibility of making ToM predictions or explanations from the starting mental states without the need for an exhaustive third person account of the causal interactions between mental states. By using starting states of the target person as inputs for their 1st person mental machinery, the simulating agent "has for themselves" the mental states of the target person, which can be read "off-line" as a prediction or explanation of what the target person will think, feel or do. But this distinctive feature of ST has no necessary relation with introspection. Indeed, Nichols and Stich (2006) argue that introspection on one's own thoughts and feelings may require quite separate cognitive apparatus from that used for ToM.

Second, ST makes no distinctive claims about the role of perceived similarity between self and other in ToM judgements. The validity of simulation does indeed rest on an assumption that the target is relevantly similar to the simulator (e.g., Heal, 1986). But this "relevant similarity" is that all humans have cognitive systems whose mental states interact according to the same basic causal principles, and this similarity is underwritten by our common biological heritage, not by a calculation of perceived similarity to self. It does not matter to ST whether the *content* of the other's mental states are similar to one's own or different: this problem is dealt with by ensuring that the simulation receives a set of inputs or "starting conditions" that are appropriate to the target other.

For example, imagine my task is to predict how a target individual will react when I tell them of my voting intentions in a forthcoming election. Calculating the target's own political

views will be an important component of supplying the appropriate starting conditions, and, indeed, may involve evaluation of the target's similarity to myself. However, whether I judge that the target's political views are similar to my own, or diametrically opposed, this information is only relevant for determining the starting conditions for the simulation. The fundamental premise of simulation accounts is that my own mind is a suitable model for the causal processes by which mental states interact. With inputs appropriately tailored to the target, my own mind should be able to predict the target's reaction to my stated voting intentions, whether the target's views are similar to mine or different.

Simiarity to self and the problem of "starting conditions" for ToM judgements. Of course, it needs to be recognised that calculation of the appropriate set of starting conditions may itself be a very complex problem. Indeed, it may only be through the lens of the ST/TT debate that this has come to be seen as a peripheral rather than a focal issue in how we understand the minds of others. Future research may well find questions about "starting conditions" for a ToM judgement to be at least as important as whether the subsequent judgement works by ST or TT. It seems likely that assessments of self-other similarity may play an important role in solving this problem by guiding the reasoner to draw analogies between themselves and the target in an appropriate way. Likewise, it will be important for future work to determine the importance of explicit, strategic processes such as introspection as well as more implicit and automatic processes of "person perception" (e.g., Macrae & Bodenhausen, 2000; Willis & Todorov, 2006). However, as is apparent in Figure 1, the problem of calculating appropriate starting conditions, and the potential role for self-other similarity in the process is just as apparent for TT as for ST. Thus, conditionalisation of neural activity involved in theory of mind judgements according to perceived similarity to self cannot provide evidence to discriminate ST from TT.

Conclusions.

Simulation-Theory and Theory-Theory currently define the terms of most philosophical debate about the nature of folk psychology, and the influence of this debate extends into the empirical literature on theory of mind. However, it has proved extremely difficult to find behavioural phenomena that provide clear evidence in favour of either ST or TT. Importantly, this may not be due to a lack of ingenuity on the part of experimental psychologists. Rather, as Stich and Nichols (1997) suggest, different versions of ST and TT may be so theoretically diverse that it is very difficult to derive distinctive behavioural predictions. Interestingly, Stich and Nichols (1997) are more optimistic about the promise of neuroimaging providing conclusive data. I have argued that existing attempts to fulfil this promise have failed, and that the prospects for success in the future are far from certain.

Critically, however, this need not reduce the contribution that social cognitive neuroscience can make to our understanding of ToM. Although the studies described above fail to provide evidence that discriminates ST from TT, they succeed in demonstrating that social cognitive neuroscience has powerful empirical and conceptual tools for investigating ToM processes. Ramnani and Miall's (2004) task successfully generated a contrast between participants making a decision themselves and participants anticipating similar decisions in other people. For the reasons already discussed, it may prove difficult to extend their design for studying behaviour governed by simple rules to behaviour governed by mental states such as beliefs and desires. However, exploring the limits to which this approach might be pushed will undoubtedly extend our understanding of social cognitive processes both with and without explicit representation of mental states. By identifying which neural systems are involved in such processes, and the degree to which they are overlapping or distinct for self and other, findings of this kind can provide valuable additions to behavioural data for motivating the development of cognitive theories. For example, if the computational processes for

appropriate self and other conditions are sufficiently similar we may find no observable differences for these processes on behavioural measures such as reaction times. Neuroimaging has the potential to reveal that processes that cannot be discriminated with behavioural measures nonetheless activate distinct neural populations. Such a finding would motivate a new hypothesis, that these processes are at least partially independent, which might be tested by examining whether 1st and 3rd person processes could be independently affected by brain damage, or trans-cranial magnetic stimulation.

Grezes et al. (2004) and Vogeley et al. (2001) illustrate successful use of comparisons between 3rd person judgements for self and other. In different ways they highlight the fact that, if common processes are used for making judgements about self and other, then we must also ask how and when a distinction between self and other is maintained (see e.g., Blakemore & Frith, 2003, for a discussion of the same issue in relation to action). In Grezes et al. (2004) similar regions of premotor cortex were activated when observing videos of both self and others lifting a box. However, these regions also discriminated between self and other by showing a more rapid response for self. This illustrates one way in which the functional and neural systems for processing actions might contain implicit information that distinguishes self from other. In this study, it is unclear whether the discrimination affected behaviour, since the accuracy with which participants detected violations in the actors' expectations did not differ whether the actor was the participant themselves or another person[2] and behavioural response times could not be measured. However, the study raises important questions about the functional relevance of such implicit discrimination between self and other, which deserve to be examined in further work.

In Vogeley et al. (2001) similar regions of right anterior cingulate cortex and left temporopolar cortex were activated for judging mental states in both self *and* other, plus unique activation was observed in right prefrontal cortex when judgements were made about

others in the same scenario that required judgements about self. Caution is clearly necessary when interpreting such findings. Since it is possible for quantitative changes in the difficulty of a given task to result in the recruitment of distinct neural systems (Stuss, Toth, Franchi et al., 1999) it could be that self and other judgements differ only in difficulty, not in kind. However, a theoretically interesting alternative is that unique neural activation for judgements about others reflects a qualitatively distinct cognitive control process involved in differentiating self from other, which may be critical for resisting interference from self-perspective when making judgements about the mental states of others (e.g., Vogeley et al., 2001; see also Decety & Grezes, 2006). This conjecture receives some support from recent neuropsychological studies, reporting a patient with a right frontal lesion whose ability to reason about other people's perspectives seems highly sensitive to interference from his own (self) perspective (Samson, Apperly, Kathirgamanathan, & Humphreys, 2005; Samson, Apperly, & Humphreys, 2007). Once again, the tools of social cognitive neuroscience are a valuable addition to behavioural experiments in suggesting how complex ToM processes might be decomposed into component parts.

Mitchell et al. (2005, 2006) have provided evidence suggesting the existence of neural systems involved differentially in making judgements about the mental states of other people depending upon whether they are perceived to be similar or dissimilar to self. As described above, this does not provide evidence of ST over TT, because similarity to self might be exploited to help provide the inputs for judgements about others on either ST or TT. However, rather than showing methodological inadequacy, this perhaps serves to illustrate how the ST/TT interpretive framework can lead to more confusion than clarity. As Stich and Nichols (1997) point out, part of the difficulty in deriving distinctive predictions for ST versus TT stems from the fact that both ST and TT reasoning requires appropriate inputs, that these inputs may themselves be the product of a ST or a TT process, and that there is no systematic basis

for drawing a line between the inputs to a particular reasoning episode and the start of the reasoning itself. Given this, it may be productive to interpret Mitchell et al.'s findings more transparently. First, they show that social judgements can be modulated by the similarity that we perceive between ourselves and the target. Second, they suggest how perceived similarity to self might achieve some of its effects, by increasing the tendency for participants to introspect and project their own views onto people they perceive to be similar to themselves, and, perhaps, by increasing participants' reliance on non-personal social semantic information for judgements about people who they perceive as dissimilar. Third, they suggest that the different strategies of introspection-projection and reasoning from social semantic information make use of dissociable functional and neural systems, located in ventral and dorsal mPFC respectively.

Such findings should make researchers interested in ToM think much more carefully about the potentially diverse processes that are likely to contribute to mental state ascription. For instance, it would be particularly interesting to know the scope of the effects of perceived similarity to self on different kinds of social judgement. Mitchell et al. (2005; 2006) had participants attribute likely thoughts or feelings to a target based on an impression formed from background information (e.g., participants saw a photograph of the target, or were told the target's political views). In contrast, most work in the ToM literature requires participants to infer a specific mental state (e.g., target thinks the object is in the red box) given a particular set of circumstances that specifically warrant that mental state (e.g., the target saw the object in the red box and did not see when the object was moved to the green box). Would the target's perceived similarity to self also affect the cognitive processes used for mental state ascription in these circumstances? How might perceived similarity to self interact with other factors such as conflict between what the target thinks and what the participant knows to be the case (as in false belief tasks)? Addressing such questions would certainly

advance our understanding of the functional architecture of ToM, whether or not the debate between ST and TT was advanced in any way.

In sum, social cognitive neuroscience has so far failed to provide evidence to distinguish between the two dominant theoretical accounts of ToM: ST and TT. However, as illustrated by the studies reviewed above, the neuroscientific approach offers unique ways of advancing our understanding of ToM processes that do not depend upon ST or TT for their validity. For example, we have a great deal more to learn about how 3rd person judgements about others relate to analogous 1st person "self" processes, about the similarities and differences between 3rd person ToM ascriptions to self and other, and about how perceived similarity to self influences the resources we draw upon when making ToM judgements. As these findings accumulate, it may still be tempting to weigh the evidence against ST and TT. The ultimate test will not be whether ST or TT can accommodate the data, but whether interpretation within the ST/TT framework yields any additional understanding of the nature of ToM processes. On the basis of the current literature it seems possible that these theories will in fact become redundant as new findings about ToM motivate the development of new models based upon well-characterised cognitive and neural processes.

References

Apperly, I.A., Samson, D., & Humphreys, G.W. (2005). Domain-specificity and theory of mind: Evaluating evidence from neuropsychology. *Trends in Cognitive Sciences, 9(12),* 572-577.

Baker, L., (1995). *Explaining Attitudes: A Practical Approach to the Mind*. Cambridge: Cambridge University Press.

Bermudez, J.L. (2005). Philosophy of psychology: A contemporary introduction. New York: Routledge.

Blakemore,S.J., Frith,C. (2003). Self-awareness and action. *Current Opinion in Neurobiology 13(2),* 219-224.

Camerer, C., Lowenstein, G., & Weber, M. (1989). The curse of knowledge in economic settings: An experimental analysis. Journal of Political Economy, 97, 1232-1254.

Carruthers, P. & Smith, P.K. (Eds.) (1995) *Theories of theories of mind.* Cambridge: Cambridge University Press.

Churchland, P.S. (1986) *Neurophilosophy: Toward a unified science of the mind/brain.* Cambridge, MA: MIT Press.

Currie, G. & Ravenscroft, I. (2002) *Recreative Minds: Imagination in Philosophy and Psychology*, Oxford University Press.

Decety, J. & Grezes, J. (2006). The power of simulation: Imaging one's own and other's behaviour. *Brain Research 1079.* 4-14.

Dennett, D.C. (1987). *The intentional stance*. Cambridge: Cambridge University Press.

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford, CA: Stanford University Press.

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

Fodor, J.A. (1975). *The language of thought*. New York: Crowell.

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

Frith, C.D., & Frith, U. (2006) The neural basis of mentalizing. *Neuron*. *50(4)* 531-534.

Gallese, V., & Goldman, A., (1998) Mirror neurons and the simulation theory of mind-reading. *Trends in Cognitive Sciences, 2,* 493-501.

Gallese, V., Keysers, C., & Rissolatti, G. (2004) A unifying view of the basis of social cognition. *Trends in Cognitive Sciences, 8(9),* 396-403.

Goldman, A.I. (1989) Interpretation psychologized. *Mind and Language 4*, 161–185.

Goldman, A.I. & Sripada, C.S., (2005) Simulationist models of face-based emotion recognition. *Cognition, 94(3)* 193-213.

Gopnik, A. & Meltzoff, A. (1997) *Words, Thoughts and Theories.* Cambridge MA: MIT Press.

Gopnik, A. & Wellman, H. (1992) Why the Child's Theory of Mind Really is a Theory. *Mind and Language 7,* 145-71.

Gordon, R. (1986) Folk psychology as simulation. (1986) *Mind and Language, 1*, 158–170.

Grezes, J., Frith, C.D. & Passingham, R.E. (2004a) Inferring false beliefs from the actions of oneself and others: an fMRI study. *Neuroimage, 21.* 744-750.

Grezes, J., Frith, C.D. & Passingham, R.E. (2004b). Brain mechanisms for inferring deceit in the actions of others. *The Journal of Neuroscience, 24(24)*. 5500-5505.

Gusnard, D. A., Akbudak, E., Shulman, G. L., & Raichle, M. E. (2001). Medial prefrontal cortex and self-referential mental activity: Relation to a default mode of brain function. *Proceedings of the National Academy of Sciences, U.S.A., 98*, 4259 – 4264.

Harris, P., (1989) *Children and Emotion*, Oxford: Blackwell Publishers.

Macrae, C. N., & Bodenhausen, G. V. (2000). Social cognition: Thinking categorically about others. *Annual Review of Psychology, 51,* 93–120.

Heal, J. (1986) Replication and functionalism. In Butterfield, J., (ed) *Language, Mind and Logic*. Cambridge: Cambridge University Press.

Milner, A.D. & Goodale, M.A. (1992). Separate visual pathways for perception and action. *Trends in Neuoscience 15(1)* 20-25.

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience,17(8)*, 1306-1315.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron, 50,* 655-663.

Nichols, S. and Stich, S. (2003) *Mindreading: An Integrated Account of Pretence, Self-Awareness, and Understanding of Other Minds*, Oxford University Press.

Nichols, S. & Stich, S. (2006, to appear). "Reading One's Own Mind: Self-Awareness and Developmental Psychology." In M. Ezcurdia, R. Stainton & C. Viger. (Eds.) New Essays in Philosophy of Language and Mind, a supplemental volume of the Canadian Journal of Philosophy.

Ramnani, N., & Miall, R.C. (2004) A system in the human brain for predicting the actions of others. *Nature Neuroscience*, 7, 85–90.

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,* 1102-1111.

Samson, D., Apperly, I.A., & Humphreys, G.W. (2007). Error analysis in brain-damaged patients with perspective taking deficits: A window to the social mind and brain. *Neuropsychologia 45(11),* 2561-2569.

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

Saxe, R. (2006) How and why to study theory of mind with fMRI. *Brain Research, 1079.* 57-65.

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

Searle, J. (1983). *Intentionality* (1st ed., Vol. 1). Cambridge: Cambridge University Press.

Spelke, E.S. & Kinzler, K.D. (2007). Core Knowledge. *Developmental Science, 10(1)*, 89-96.

Stich, S. & Nichols, S., (1992) Folk psychology: Simulation or tacit theory? *Mind and Language, 7*, 35-71

Stich, S., & Nichols, S. (1997). Cognitive penetrability, rationality, and restricted simulation. *Mind and Language, 12,* 297-326.

Stuss, D.T., Toth, J.P., Franchi, D., Alexander, M.P., Tipper, S., & Craik, F.I.M. (1999) Dissociation of attentional processes in patients with focal frontal and posterior lesions. *Neuropsychologia* 37, 1005-1027

Vogeley, K., Bussfield, P., Newen, A., Herrmann, S., Happe, F., Falkai, P., Maier, W., Shah, N.J., Fink, G.R. & Zilles, K. (2001) Mind reading: Neural mechanisms of theory of mind and self-perspective. *Neuroimage 14.* 170-181.

Vogeley, K., May, M., Ritzl, A., Falkai, P., Zilles, K., & Fink, G. R. (2004). Neural correlates of first-person perspective as one constituent of human self-consciousness. *Journal of Cognitive Neuroscience, 16,* 817– 827.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after 100 ms exposure to a face. *Psychological Science, 17(7),* 592-598 .

Footnotes

1. My focus in the current article will be upon propositional attitudes, of which “belief” is the paradigm case. The ability to ascribe propositional attitudes such as beliefs is (rightly or wrongly) the central focus in empirical research on ToM, and theoretical research on ST and TT. It is important to recognise that other abilities falling under a broader definition of ToM, such as ascribing basic emotional states or perceiving action, have also been evaluated in terms of the debate between ST and TT (e.g., Decety & Grezes, 2006; Goldman & Sripada, 2004). In particular, research on mirror neurons has been seen by many to suggest that simulation is a common social-cognitive process (e.g., Gallese & Goldman, 1998; Gallese, Keysers & Rizzolati, 2004). However, these researchers typically stress that mirror neurons could not explain the ability to ascribe propositional attitudes such as beliefs, even if they are an important phylogenetic or ontogenetic precursor to such abilities (e.g., Decety & Grezes, 2006; Gallese & Goldman, 1998). Thus, the literature on mirror neurons and emotion ascription is not thought to provide direct evidence about the role of ST or TT in the ascription of propositional attitudes, and will not be discussed in the current paper.

2. There was in fact a non-significant trend for less accurate judgements about self that may have been significant in a more substantial sample.

3. As already noted (in Footnote 1), a much stronger case can be made that simulation has a role in ascribing emotions. However, emotions are not propositional attitudes. They have not been the main focus for the ST/TT debate or of this paper.

Acknowledgements

I would like to thank Sarah Beck, Steven Butterfill, Kimberly Quinn, Dana Samson, Rebecca Saxe and two anonymous reviewers for their comments on drafts of this manuscript.