What Functional Magnetic Resonance Imaging can tell us about Theory of Mind

by

Rebecca R Saxe

BA (Hons) Psychology and Philosophy
University of Oxford, 2000

SUBMITTED TO THE DEPARTMENT OF BRAIN AND COGNITIVE SCIENCES IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF

PH.D. IN COGNITIVE SCIENCE
AT THE
MASSACHUSSETTS INSTITUTE OF TECHNOLOGY
2003
[September 2003]

Signature of Author:

Rebecca Saxe

Certified by:

Nancy Kanwisher
Professor of Cognitive Neuroscience.

Accepted by:

Earl Miller
Picower Professor of Neuroscience
Chairman, Department Graduate Committee

What Functional Magnetic Resonance Imaging can tell us about Theory of Mind

by

Rebecca R Saxe

Submitted to the Department of Brain and Cognitive Sciences
In partial fulfilment of the requirements for the degree of
Ph.D in Cognitive Science

Abstract

To have a theory of mind is to be able to explain and predict human behaviours and experiences in terms of mental states: beliefs, desires, goals, thoughts, and feelings. In chapters 1 and 2, I use functional magentic resonance imaging (fMRI) to investigate the neural substrate of the theory of mind, in healthy human adults. I conclude (1) that specialised brain regions, including a region of the temporo-parietal junction (the TPJ-M), are selectively engaged when people reason about the contents of other people's beliefs, and (2) that the brain regions associated with belief attribution appear to be distinct from other regions engaged in the representation of goal-directed action, including a region of posterior superior temporal sulcus (the pSTS-VA). In chapters 3 and 4, I consider the implications of these and other neuroimaging results for the mental structure of theory of mind, based on proposals derived from developmental psychology and philosophy.

Thesis Supervisor: Nancy Kanwisher
Professor of Cognitive Neuroscience

Acknowledgements

Nothing in this dissertation would have been possible without my primary advisor, Nancy Kanwisher, and it is dedicated to her. I thank her for her vehement faith in me, for her unfailing enthusiasm and for her truly extraordinary generosity.

I have always been incredibly lucky in my advisors, and I am deeply grateful to them all. In addition to Nancy, Kia Nobre, David Perrett, Josh Tenenbaum and Susan Carey have been exemplary mentors, teachers and colleagues. These are my giants.

A thousand thanks to my committee who provided enough impetus, flexibility and encouragement to allow me to finish this thesis on short notice. I have been particularly fortunate in the wise guidance and continued support of Steven Pinker.

Yuhong Jiang and Andrea Heberlein have been good friends as well as good editors throughout this process. Chris Baker, Tania Lombrozo, and Josh Greene also edited and considerably improved parts of the manuscript. David Badre provided many conversations about science and life.

A crew of fabulous undergraduates helped with the collection and analysis of the data in chapters 1 and 2: Ben Balas, Robb Rutledge, Amal Dorai, Steve Lee and Christine Wang. John Rubin, Dengke Xiao and Gyula Kovacs helped make the movie stimuli. Miles Shuman and Nick Knouf provided invaluable help with the fMRI.

Finally, I want to thank my family and friends, who have remained firm in their convictions about me, and especially my wonderful parents, Dianne and Stewart Saxe, and my husband, Jonah Steinberg.

Chapter 1 is reprinted with permission from Neuroimage, Volume 19 © 2003 by Elsevier Science.

Chapter 3 is reprinted, with permission, from the Annual Review of Psychology, Volume 55, ©2004 by Annual Reviews www.annualreviews.org.

For Nancy, with thanks.

Introduction

This dissertation is about the human 'theory of mind', our faculty for thinking about thoughts. "What does he want?", we ask ourselves. "How will she feel tomorrow?" And perhaps most commonly, "What was I thinking?" In our theory of mind lie the conceptual resources to ask these questions, and often to answer them.

There are four fundamental questions to ask about any human psychological function (adapted from Chomsky 1988). Applied to this domain, they become:

1. What is a theory of mind? What is in the mind of a healthy human adult that lets her reason about the workings of her own and other minds?
2. How does a theory of mind arise? How do children develop competence in psychological reasoning?
3. How is a theory of mind actually used in the explanation of particular behaviours? In the prediction/anticipation of future behaviour?
4. What are the physical mechanisms in the human brain that serve as the basis for a theory of mind?

A full account of the human theory of mind would provide mutually compatible and sufficient answers to all of these questions. In this dissertation I propose a more modest project. The first two chapters are addressed particularly to question (4), how the theory of mind is implemented in the human brain, using functional magnetic resonance imaging (fMRI). Then in the latter two chapters, I widen the lens a little, to ask how the answers to question (4) suggested by neuroimaging results can influence broader theoretical debates about the structure, development and application of theory of mind (questions 1, 2 and 3, in chapters 4, 3, and 4 respectively).

The very beginning of this discussion must be to provide a reference for the central term, 'theory of mind'. I will use 'theory of mind' as I believe it was originally intended, to refer to the psychological faculty or faculties that allow all healthy human adults to impute unobservable mental states to themselves and others, and to use these attributed mental states as a coherent framework to make remarkably successful inferences about the contents of other mental states, and predictions and explanations of human behaviour. An important principle in generating these inferences is the so-called Principle of Rationality: we expect that in general a person will act to achieve her desires, given her beliefs. Adults also understand that while beliefs ought to represent the world correctly, they sometimes fail to do so, and when this occurs the person's subsequent behaviour will be guided by the contents of false beliefs.

How is this theory of mind implemented in the human brain? A first approach to this question is to ask whether the brain contains specialised mechanisms for theory of mind (as do vision and language) or whether theory of mind inferences depend on the same brain mechanisms as other difficult cognitive problems like general relativity theory or chess. My results support the former, "domain specific" answer. The studies reported in chapter 1 show that a region in the human temporo-parietal junction (here called the TPJ-M) is involved specifically in reasoning about the contents of mental states. The TPJ-M response is high during the attribution of both true and false beliefs, but not during (a) any other reasoning about people or (b) reasoning about non-mental representations (e.g. photographs) or (c) hidden causes in general.

The evidence in Chapter 1 for a domain specific neural system for theory of mind in healthy human adults is also consistent with previous evidence from people with Autism (and the related Asperger's Syndrome). Autism is characterised by social, linguistic and general cognitive disabilities, included marked deficits in pretend play, social communication and forming affective relationships (Kanner 1943). Children with autism have particular difficulty on tasks that require them to predict and explain behaviour based on the contents of a person's beliefs and desires. But when the same prediction task is rephrased so that it refers only to non-mental representations like photographs the autistic children succeed, outperforming the normal controls (Leslie and Thaiss 1992). The selective impairment in Autism for inferences about mental states is consistent with specialised neural mechanisms that apply only in this domain, like the TPJ-M.

In Chapter 2, I elaborate the characterisation of the specialised neural system for reasoning about other people, showing that it has at least three components. Neighbouring the TPJ-M but distinct from it are two other brain regions: the Extrastriate Body Area (EBA), involved in representing human body form (and possibly motion, Downing et al 2001, Grossman et al 2001), and a region in posterior Superior Temporal Sulcus involved in the Visual analysis of Action (pSTS-VA). In particular, abstract visual descriptions of actions that permit interpretation of intentions appear to be the domain of the pSTS-VA, whether or not they are defined by human bodies moving in a characteristic articulated fashion. I conclude that the vicinity of the superior temporal sulcus contains neural mechanisms for representing bodies (the EBA), actions (the pSTS-VA) and thoughts (the TPJ-M).

The division of labour between the TPJ-M and the pSTS-VA is akin to the distinction between two components of the developing theory of mind: an early-developing system for reasoning about goals, perceptions, and emotions, and a later-developing system for representing the contents of beliefs. In Chapter 3, I consider in

depth the correspondence between the two components of theory of mind suggested by developmental psychology and neurimaging. To the extent that this correspondence holds, the neuroimaging results in Chapter 2 suggest that neural mechanisms related to both the earlier and the later systems of theory of mind remain intact and distinct in adult brains.

A two-component neural system for theory of mind has also been predicted from neuropsychological evidence. Tager-Flusberg and Sullivan (2000) reported that a group of children with William's Syndrome (WS), a disorder caused by a mutation of a single gene, shared only one component of the deficit of theory of mind observed in children with Autism. Like the autistic children, WS children failed to predict behaviour based on the contents of mental states. However on a different task, requiring more 'perceptual' judgements about mental states (which of two mental states terms best described a photo of the eye region of a human face), WS children succeeded while Autistic children continued to fail. Tager-Flusberg and Sullivan (2000) argued that this between-groups dissociation of deficits was evidence for two distinct components of reasoning about others: a "social-cognitive" component impaired in both groups and a "social-perceptual" component impaired only in Autistic people. These components may also correspond to the later-developing component of theory of mind (and the TPJ-M), and the earlier-developing component (and the pSTS-VA), respectively.

In all, the data in chapters 1 and 2 provide a preliminary sketch of the place of theory of mind in the adult human brain. Theory of mind function is supported by specialised neural mechanisms, with at least two distinct components. These components are associated with distinct brain regions including the pSTS-VA for representing goal-directed action and the TPJ-M for representing the contents of mental states. These conclusions are a far cry from a complete account of how theory of mind is neurally implemented. Nevertheless even these rough approximations have implications for the other fundamental questions about theory of mind, and so I turn to those questions in Chapters 3 and 4.

Chapter 3 considers the implications of the preceding (and other) neuroimaging results for theories of theory of mind development. The continued existence of distinct and specialised neural mechanisms in adults corresponding to the two primary components of the developing theory of mind suggests that that the two developmental stages reflect the emergence of two distinct systems, rather than the elaboration or unmasking of a single system. Importantly, the brain regions associated with the later-developing system for theory of mind are distinct from brain regions engaged in inhibitory control and in syntactic processing. The development of both inhibitory control and complement structure syntax are correlated with the emergence of belief attribution,

raising concerns that belief attribution skills were merely being unmasked by competence in either or both of these other domains. The clear neural distinction between these processes is evidence to the contrary, that belief attribution is not dependent on either inhibitory control or syntax at least in adulthood, but is subserved by a specialised neural system for theory of mind.

Finally, in chapter 4, I turn to the mental structure of theory of mind, and Chomsky's first question: what is a theory of mind? I consider three candidate accounts: modules, theories and off-line simulation. The discussion also touches on an answer to Chomsky's third question: how a theory of mind could applied, to make actual inferences about behaviour and mental experiences. For these most fundamental questions, though, the contribution of neuroimaging is relatively limited, and awaits future work.

References

Chomsky N. 1988. Language and Problems of Knowledge. Cambridge MA: MIT Press.

Downing PE, Jiang Y, Shuman M, Kanwisher N. 2001. A cortical area selective for visual processing of the human body. *Science* 293: 2470-3

Grossman, ED, Blake R. 2001. "Brain activity evoked by inverted and imagined biological motion." Vision Res **41**(10-11): 1475-82.

Kanner L (1943) Autistic disturbances of affective contact. Nervous Child 2:217-250.

Leslie A, Thaiss L. 1992. Domain specificity in conceptual development. *Cognition* 43: 225-51 Malle BF. 2001. Folk explanations of intentional action. In *Intentions and Intentionality: Foundations of Social Cognition*, ed. BF Malle, LJ Moses, DA Baldwin. Cambridge MA: The MIT Press

Tager-Flusberg H, Sullivan K. 2000. A componential view of theory of mind: evidence from Williams syndrome. *Cognition* 76: 59-90.

Chapter 1. People thinking about thinking people: the role of the temporo-parietal junction in Theory of Mind.

The remarkable human facility with social cognition depends on a fundamental ability to reason about other people. Specifically, we predict and interpret the behaviour of people based on an understanding of their minds: that is, we use a 'Theory of Mind'[1]. In this paper we show that a region of human temporo-parietal junction is selectively involved in reasoning about the contents of other people's minds.

Brain regions near the temporo-parietal junction (TPJ) have been implicated in a broad range of social cognition tasks (Allison, Puce and McCarthy 2000, Gallagher and Frith 2003, Green and Haidt 2003). Regions near the TPJ have preferential responses to human faces (e.g. Hoffman and Haxby 2000), bodies (e.g. Downing, Jiang, Shuman and Kanwisher 2001) and biological motion (e.g. Grossman et al 2000). There is also some evidence that regions within human TPJ are involved in Theory of Mind (ToM). A number of studies have reported increased responses in the TPJ when subjects read verbal stories or see pictorial cartoons that require inferences about a character's (false) beliefs, compared with physical control stimuli (Fletcher, Happe et al 1995, Brunet, Sarfati, Hardy-Bayle and Decety 1998, Gallagher, Happe et al 2000, Castelli, Happe, Frith and Frith 2000, Voegely, Bussfeld et al 2001. A number of other brain regions have also been implicated in theory of mind; see reviews by Gallagher and Frith 2003 and Greene and Haidt 2003).

What is the role of the temporo-parietal junction in these tasks? Theory of Mind (ToM) reasoning depends upon at least two kinds of representation: a representation of *another person* per se and a representation of that other person's *mental states* (see Leslie 1999). While a representation of a person per se is a likely prerequisite for ToM, achieving a representation of others' mental states is the core responsibility of a Theory of Mind. Some authors suggest that the TPJ is involved only in the preliminary stages of social cognition that "aid" ToM, not in Theory of Mind reasoning itself (e.g. Gallagher and Frith 2003). We here provide evidence against this suggestion, and argue on the contrary that a region of the TPJ is selectively involved in representation other people's mental states.

---

[1] The term "Theory of Mind" has a more restricted sense, referring to the suggestion that the structure of knowledge in the mind is analogous to a scientific theory (e.g. Carey 1985, Wellman and Gelman 1992). For discussions about the so-called Theory-theory, see Carruthers and Smith 1996 and Malle, Moses and Baldwin 2001. In this paper, we use the term "Theory of Mind" in a broader sense, to refer to any reasoning about another person's represenational mental states (also called 'Belief-Desire psychology', e.g. Bartsch and Wellman 1995).

Neuroimaging studies have followed developmental psychology in using "false belief" stories as the prototypical problem for Theory of Mind reasoning (Fletcher et al, 1995; Gallagher et al, 2000, see also Vogeley et al 2001). In these scenarios, a character's action is based on the character's false belief (Wimmer and Perner 1983). False beliefs provide a useful behavioural test of a ToM, because when the belief is false, the action predicted by the belief is different from the action that would be predicted by the true state of affairs (Dennett 1978). Note, though, that everyday reasoning about other minds, by adults and children, depends on attributions of mostly true beliefs (e.g. Dennett 1996, Bartsch and Wellman 1995).

Previous investigations of the neural correlates of Theory of Mind (Fletcher et al, 1995, Gallagher et al, 2000) have compared False Belief ("Theory of Mind") stories with two control conditions: "Non Theory of Mind Stories", which describe actions based the character's true beliefs, and "Control" stories, consisting of unrelated sentences. These authors found that the TPJ response was high during "Theory of Mind" stories, but was also high during "Non Theory of Mind" stories. They concluded (see also Gallagher and Frith 2003) that the TPJ is not selectively involved in Theory of Mind. This conclusion does not follow. Because the "Non Theory of Mind" stories invite inferences about the character's (true) beliefs, a region involved in reasoning about other minds *should* show a high response to these stories, as well as to the so-called "Theory of Mind" stories. (For an argument against the use of unrelated sentences as the baseline condition, see Ferstl and von Cramon 2002).

We propose two basic tests for a region selectively involved in Theory of Mind reasoning. First, it must show increased response to tasks/stimuli that invite Theory of Mind reasoning (about true or false beliefs) compared with logically similar non-social controls. Second, the region must respond not just when a person is present in the stimulus, but specifically when subjects reason about the person's mental states. Below, we provide evidence that a sub-region of the TPJ, here called the TPJ-M, passes both these criteria for a selective role in Theory of Mind.

Experiment 1

We devised a new version of the false belief stories task (Fletcher, Happe et al 1995)to compare reasoning about true and false beliefs to reasoning about non-social control situations. **ToM** stories described a character's action caused by his/her false belief. Descriptions of **Human Actions** required analysis of mental causes, in the absence of false beliefs. We compared these conditions to two non-social control conditions: **Mechanical Inference** control stories, which required the subject to infer a hidden

physical (as opposed to mental) process, such as melting or rusting (for examples, see Appendix 1), and **Descriptions of Non-Human objects**.

Unlike previous studies, we did not cue or instruct subjects to attend specifically to mental states. With this design we were able to look for regions of cortex in individual subjects that are selectively and spontaneously involved in understanding the mental (as opposed to physical) causes of events.

To test whether the response to ToM stories was a response to the presence of a person in the stimulus, we presented still photographs of **People**, and non-human **Objects**. Downing et al (2000) reported a bilateral region near the posterior superior temporal cortex that responds preferentially to the visual appearance of human bodies, compared with a range of control objects (the Extrastriate Body Area, EBA). We tested directly the functional and anatomical relationship between the EBA and the (proposed) TPJ-M.

Methods

Twenty-five healthy right-handed adults (12 women) volunteered or participated for payment. All subjects had normal or corrected to normal vision and gave informed consent to participate in the study.

Subjects were scanned in the Siemens 1.5 (9 subjects) and 3.0 T (16 subjects) scanners at the MGH-NMR center in Charlestown, MA, using a head coil. Standard echoplanar imaging procedures were used (TR = 2 sec, TE = 40 [3T] or 30 [1.5T] msec, flip angle 90°). Twenty 5mm thick near-coronal slices (parallel to the brainstem) covered the occipital lobe and the posterior portion of the temporal and parietal lobes.

Stimuli consisted of short center-justified stories, presented in 24-point white text on a black background (average number of words = 36). Stories were constructed to fit four categories: False Belief, Mechanical Inference, Human Action and Non-Human Descriptions (Appendix 1). Each story was presented for 9500 ms, followed by a 500 ms interstimulus interval. Each scan lasted 260 seconds: four 40-second epochs, each containing four stories (one from each condition), and 20 seconds of fixation between epochs. The order of conditions was counterbalanced within and across runs. Subjects were asked to press a button to indicate when they had finished reading each story. Subjects read a total of 8 (4 subjects) or 12 (21 subjects) stories per condition.

Fourteen of the subjects from Experiment 1 (7 women) were also scanned on an EBA localiser in the same scan session, all at 3.0 T. Stimuli consisted of twenty grayscale photographs of whole human bodies (including faces) in a range of postures, standing and sitting, and twenty photographs of easily recognisable inanimate objects (e.g. car, drum,

tulip). (Two other conditions, cropped faces and scrambled objects, were included in the scan but were not analysed here).

Image presentation followed the blocked design described in Tong, Nakayama, Moscovitch, Weinrib, & Kanwisher (2000, Experiment 1) except that images were presented at a rate of one every 800 ms (stimulus duration = 500 ms, interstimulus interval = 300 ms), and each scan lasted 336 seconds. Subjects performed a one-back matching task (Tong et al 2000).

MRI data were analysed using SPM 99, FS-fast and in-house software.

Results

Average reading times for Theory of Mind and Mechanical Inference stories did not differ significantly (ToM = 6.4 s, MI = 6.5 s, $p>0.2$).

Random effects analyses of 25 subjects revealed five loci of greater activation during the theory of mind compared with mechanical inference stories ($p<0.05$ corrected for multiple spatial hypotheses): left and right temporo-parietal junction (TPJ-M), left and right anterior superior temporal sulcus (aSTS), and precuneus (Table 1, Figure 1.1). (Consistent with many previous studies (e.g. Raichle et al 2000) the precuneus was deactivated [BOLD signal less than fixation baseline] during all of our story conditions. The ToM stories deactivated the precuneus less than Mechanical Inference stories. It was therefore unclear whether this effect should be considered a response to ToM or to Mechanical Inference stories, and the precuneus response was not analysed further).

The same pattern of results was apparent in individual subjects (Fixed Effects $p<0.0001$ uncorrected for all results reported here). Voxels more responsive during Theory of Mind than Mechanical Inference stories were observed at the TPJ in 22 out of 25 subjects (bilaterally in 14, left in 5, and right in 3 subjects). The aSTS activation was significant at this level in 10 out of 25 subjects. Because the TPJ-M was most consistent across subjects and was the focus of our prior hypotheses, we concentrated on this region in the subsequent analyses.

We defined TPJ-M regions of interest (ROI) in the left and right TPJ in each individual subject as contiguous voxels in each hemisphere that were more active ($p<0.0001$) during False Belief than Mechanical Inference stories. The TPJ-M bilaterally generalised beyond false beliefs, responding significantly more to Human Action (HA) stories than to Non-human Descriptions (N-H D, paired-samples t-tests, right: HA average percent signal change from fixation (PSC): 0.22, N-H D average PSC: 0.02, $p<0.0001$; left: HA average PSC: 0.35, N-H D average PSC: 0.10, $p<0.0001$).

In the fourteen subjects who also had an EBA localiser scan, EBA ROIs were defined as the cluster of contiguous voxels in extrastriate cortex (bilaterally in 13 subjects and right-only in one subject) that was more active ($p<0.0001$) during pictures of human bodies than during pictures of non-human objects in each individual subject (following Downing et al 2000). Both right and left EBA ROIs failed to discriminate between any story conditions (paired samples t-tests, all $p>0.4$, all story PSCs below 0.01, Figure 1.2).

TPJ-M response to photographs was lateralised. The left TPJ-M did not discriminate between photographs of People (PSC: -0.04) and of Objects (PSC: -0.09, paired samples t-test, $p>0.4$). The right TPJ-M showed a trend towards a greater response to photographs of People (PSC: 0.24) than of Objects (PSC: 0.10, paired samples t-test, $p<0.10$). A repeated-measures ANOVA of Content (Person versus Object) by Stimulus modality (Stories versus Photograph) by Hemisphere (right versus left) revealed a main effect of Person > Object ($p<0.001$) and of Stories > Photographs ($p<0.05$) modulated by an interaction between Stimulus Modality and Hemisphere (response to photographs only on the right, $p<0.005$) and a trend towards a three-way interaction (the right TPJ-M response distinguishes photographs of bodies and objects more than the left TPJ-M, $p<0.1$, Figure 1.2).

Discussion

Experiment 1 thus shows an increased BOLD response in a region of the temporo-parietal junction bilaterally, here called the TPJ-M, during Theory of Mind compared with Mechanical Inference stories. This activation is robust and reliable across individual subjects. This finding replicates the earlier reports with a new set of stimuli, a less biased task (no cues), and with more stringent statistical tests (both individual subject analyses and Random Effects group analyses). Our results confirm that the TPJ-M response to verbal descriptions generalises to human actions based on true beliefs.

Importantly, we distinguished the TPJ-M from its neighbour, the EBA, which did not respond to any verbal story conditions. However, the TPJ-M response to non-verbal social stimuli appeared to be lateralised. The left TPJ-M response was selective for verbal descriptions, while the right TPJ-M activation may generalise to non-verbal stimuli, such as photographs.

Experiment 2

The results of Experiment 1 established that bilateral regions near the temporo-parietal junction show a greater increase in BOLD signal when subjects reason about others' mental states, than when they reason about non-human objects. However in Experiment 1 stories involving people AND mental states were compared with stories

that involve neither people NOR mental states. In Experiment 2, we asked which of these two components was responsible for the observed activation. We directly compared the response of the TPJ-M to stories about people that did (Desires) or did not (Physical People) require inferences based on mental states.

Also, while they were controlled for difficulty and causal structure, the logical structure of the Theory of Mind stories used in Experiment 1 (and previous studies) differed systematically from the control stories: only the false belief stories require the notion of a false representation, in this case a false belief. This confounding factor was perceived by developmental psychologists, who invented its solution: "false photograph" stories (Zaitchik 1990) which require subjects to represent the (false) content of a physical representation such as a photograph or map.

For Experiment 2 we therefore created five new sets of stories (for examples see Appendix 2). (1) **False Belief** stories, (2) **False Photograph** stories, (3) **Desires**, (4) **Inanimate Descriptions** and (5) **Physical People**. Desire stories described a character's goals or intentions and thus rely on ToM. Non-human Description stories consisted of short descriptions of non-human objects such as plants, cars or planets. Physical People stories were short descriptions of people from a purely physical perspective: clothing, hair colour, facial markings, etc.

We predicted that regions specifically involved in Theory of Mind should have a equally low response in the Non-human Description and (critically) Physical People conditions, and a higher BOLD response in the Desire condition. By contrast, regions involved in processing any other representation of other people would show a high BOLD signal for the Physical People condition.

Methods

Twenty-one naïve right-handed subjects (11 women) were scanned at 1.5 T, using twenty 5mm thick axial slices that covered the whole brain. An additional seven subjects from Experiment 1 (4 women) also participated in part of Experiment 2. All were scanned at 3.0 T using twenty 5mm thick near-coronal slices (parallel to the brainstem) covering most of the occipital lobe and the posterior portion of the temporal and parietal lobes.

Story stimuli consisted of 70 stories (12 each of 'False Belief', 'False Photograph', 'Desire', 'Physical Description' and 'Non-human Description', average number of words = 32, see Appendix 2). After each story a two-alternative forced choice 'fill-in-the-blank' question was presented for 4 s. The question consisted of a single sentence with a word missing, presented above two alternative completions on the left

and right side of the screen. Subjects pressed the left-hand response button if the word on the left completed the sentence to fit the story, and the right-hand button to choose the word on the right. 50% of the False Belief, False Photograph and Desire story questions probed the character's mental states; the other 50% probed the actual outcome, to prevent formulaic response preparation. Subjects were given three practice trials before going into the scanner: two False Belief trials, and one False Photograph trial.

Fourteen subjects (including the seven from Experiment 1) were tested on only 'False Belief' and 'False Photograph' stories. For these subjects, each run lasted 204 seconds, and consisted of six blocks (each containing 1 story [10 s] and 1 question [4 s]), alternating between the two conditions; there were three blocks per condition per run. The remaining fourteen subjects were tested on all five conditions. Each run lasted 272 seconds, and consisted of 10 blocks (each containing 1 story [10 s] and 1 question [4 s]). There were two blocks per condition per run.

Fixations of 12 seconds were interleaved between blocks. The order of conditions was counterbalanced across runs. Behavioural data were collected during the scan.

Results

Subjects were slower when responding to questions about False Photograph than False Belief stories (FB : 2.6 vs FP: 2.8 seconds, $p<0.01$), making it unlikely that False Belief inferences were simply more difficult.

As predicted, a random effects analysis on the twenty-one subjects who underwent whole brain scanning revealed regions of increased BOLD signal to False Belief compared with False Photograph stories ($p<0.0001$ uncorrected) at the temporo-parietal junction bilaterally (right: [54 –51 **18**], left: [-48 –63 33]), precuneus/posterior cingulate ([3 –54 30]), right anterior superior temporal sulcus ([54 –18 –15]), and in medial superior frontal gyrus ([6 57 18]) in the frontal pole (Figure 1.2). Medial prefrontal cortex has repeatedly been implicated in Theory of Mind processing, both in neuroimaging and in lesion studies (e.g. Rowe, Bullock, Polkey and Morris 2001, Stuss, Gallup and Alexander 2001).

For the seven subjects who were scanned in both Experiments 1 and 2, two additional analyses were conducted to confirm that the TPJ-M was consistent across Experiments. First, in all seven subjects the TPJ-M defined in Experiment 1 overlapped strikingly with TPJ-M defined by the contrast of False Belief (FB) versus False Photograph (FP) stories in Experiment 2. Figure 1.3a shows the overlap in a typical individual subject of the TPJ-M defined by these two tasks. Second, this overlap was confirmed with a functional region of interest (ROI) analysis. Voxels near the TPJ more

active during ToM than Mechanical Inference stories in these individual subjects in Experiment 1 ($p<0.0001$ uncorrected) were probed for their response during Experiment 2. This independent ROI showed a much greater response to False Belief than False Photograph stories in Experiment 2 (mean FB PSC= 1.6, mean FP PSC= 0.7, t-test $p<0.02$, Figure 1.3b). The reliability of the TPJ-M across experiments makes it unlike that the results of Experiment 1 were the result of stimulus confounds or logical differences between conditions.

For the fourteen subjects who saw all five conditions, the fMRI data were further analysed within individually defined functional regions of interest (ROI) that included all voxels that met two criteria: they were significantly more active in at least half of the individual subjects during False Belief than False Photograph stories ($p<0.0001$ uncorrected) and they fell within a sphere of 15mm radius centred on the most significant voxel of clusters identified in the Random Effects group analysis ($p<0.0001$ uncorrected) of the same contrast. Using these criteria, we identified ROIs in the TPJ-M bilaterally and right aSTS.

In the TPJ-M and the right aSTS, the BOLD signal change during Desire stories was significantly greater than during either Physical People or Non-human Description stories (both paired samples t-tests $p<0.05$), which did not differ from each other (Figure 1.4). The left and right TPJ-M did not differ. Thus these regions are not involved in the detection of any person in verbal stories, but respond selectively to stories in which describe (or imply) character's mental states. Did any regions show the predicted profile of a response to a person per se? At a lower threshold, a separate whole brain analysis ($p<0.001$ uncorrected) of Physical People > Non-human Descriptions revealed regions of frontal cortex (dorsal medial prefrontal [-3 57 39], and right lateral frontal cortex [39 15 54]).

Discussion

The results of Experiment 2 confirm that the TPJ-M shows an increased response to stimuli that invite Theory of Mind reasoning compared with logically similar non-social controls (False Photograph stories). Second, the TPJ-M does not show an increased response to the mere presence of a person in the stimulus (Physical People stories). The right and left TPJ-M responses to Physical People stories did not differ, thus resolving the ambiguity of the apparently lateralised response to photographs of bodies in Experiment 1.

General Discussion

In two Experiments we found greater BOLD response in a region within the temporo-parietal junction bilaterally (here called TPJ-M) while subjects read stories that describe or imply a character's goals and beliefs than during stories about non-human objects. This pattern is robust across subjects, tasks, and stimuli, and is not merely an effect of the difficulty or logical structure of False Belief stories, since the TPJ-M did not respond to the more difficult and logically similar False Photograph stories.

We asked whether the TPJ-M represents the simple presence of another person (possibly via detecting a human body and/or biological motion) or is involved specifically in Theory of Mind. We found that the TPJ-M was anatomically and functionally distinct from the nearby EBA (Downing et al 2000), which responded preferentially to the visual appearance of human bodies, suggesting the presence of at least two distinct regions involved in social information processing.

A key innovation of this paper over previous studies was the inclusion in Experiment 2 of Physical People stories, which described the physical appearance of human bodies. Previous studies (Fletcher et al 1995, Gallagher et al 2000) have included 'physical' stories describing acting people, which produced greater activation in the TPJ than a scrambled sentence control. Our data show that the TPJ-M response was *no greater* to stories that described other people in physical detail than that to stories describing the physical details of non-human objects – and was significantly lower than to stories that did invite a mental state interpretation (Desire stories).

Could the TPJ-M activation reflect mental imagery of the biological motion or goal-directed action described in the False Belief, Human Action, and Desire stories? We think this is unlikely. Saxe, Xiao, Kovacs, Perrett and Kanwisher (submitted) found that the TPJ-M response to a movie of a walking person was much lower than its response to False Belief stories. If the response of the TPJ-M to verbal stories was merely a consequence of subjects' imagining biological motion, we would predict the opposite. Also the TPJ-M was doubly dissociated from its neighbour, the pSTS-VA (Visual analysis of Action), which responded more to the movies than to verbal stories.

In all, our results show that a region of the TPJ[2] is involved in reasoning about other minds, not just in understanding stories involving people per se (Gallagher and

[2] What is the relationship between the TPJ-M and attention? Selective attention leads to *increases* in regions of the TPJ during social perception tasks (e.g. Narumoto, Okada, Sadato, Fukui and Yonekura 2001, Winston, Strange, O'Doherty and Dolan 2002), and to *decreases* in regions of the TPJ during visual attention tasks (Shulman et al 1997, Gusnard and Raichle 2001, Jiang and Kanwisher, unpublished data).

Frith 2003, p.80). But critically, neighbouring sub-regions of cortex have different functional profiles, highlighting the necessity of careful within-subject comparisons. The TPJ-M, identified here by responses to (False) Belief stories, may play a broad role in social and even moral cognition (Moll J, Oliveira-Souza R et al 2002, Greene and Haidt 2003).

Downar, Crawley, Mikulis and Davis (2001) proposed "a role for the TPJ in detecting behaviourally relevant events in the sensory environment" (p. 1256) that is interfered with by demanding visual attention. One possibility is that the mental states of other people constitute a particular category of such "behaviourally relevant" stimuli. Alternatively, these results may reflect functionally and anatomically distinct sub-regions within the TPJ. Direct testing of the relationship between the TPJ-M and selective attention is an important avenue for future work.

Table 1

Region	MNI co-ordinate (max voxel)	Z	# voxels ($p<0.05$ corrected)
L TPJ-M	[-54 –60 21]	5.88	63
L aSTS	[-57 –27 –12]	5.40	55
Prec	[-9 –51 33]	5.20	41
R TPJ-M	[51 –54 27]	5.10	10
R aSTS	[66 -18 -15]	4.91	2

Experiment 1. Five regions showed increased signal during Theory of Mind, compared with Mechanical Inference, stories (Random Effects, n = 25, $p<0.05$): left and right temporo-parietal junction (TPJ-M), left and right anterior superior temporal sulcus (aSTS), and precuneus (Prec).

Figure Legends

1. Experiment 1. Random effects analysis, $p<0.05$ corrected, n = 25. Theory of Mind > Mechanical Inference stories. Cross-hair marks the most significant voxel in the left TPJ-M (1). Also visible are activations in right TPJ-M (2), left aSTS (3) and precuneus (4).

2. Experiment 1. Average percent signal change from fixation in (a) left and right TPJ-M and (b) left and right EBA, defined in individual subjects (n=14). The EBA consisted of contiguous voxels in bilateral extrastriate cortex that responded significantly more to pictures of human Bodies than pictures of non-human Objects ($p<0.0001$ uncorrected). The TPJ-M consisted of contiguous voxels near the temporo-parietal junction that responded significantly more to Theory of Mind (ToM) stories than to Mechanical Inference (MI) stories ($p<0.0001$ uncorrected). (Response magnitudes for the conditions that were used to define the ROIs are illustrative only). The EBA did not respond to story stimuli. The right TPJ-M differentiated between pictures of Bodies and of Objects ($p<0.05$, paired samples t-test) but the left TPJ-M did not. ToM = Theory of Mind (false belief) stories. MI = Mechanical Inference stories. Body = photographs of human Bodies. Obj = photographs of non-human Objects.

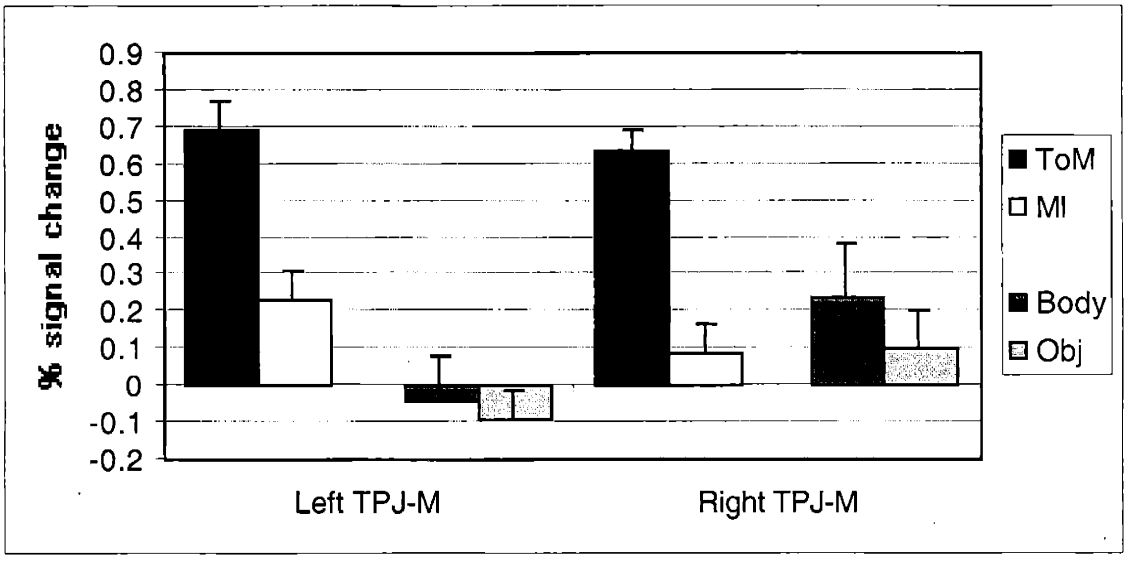
3. (a) Experiments 1 and 2. Activation overlap within an individual subject showing bilateral TPJ and precuneus regions (Fixed effects, $p<0.001$). Red = Theory of Mind > Mechanical Inference (Ex 1). Blue = False Belief > False Photo (Ex 2). Green = Both. (b) Single subject time-course of response during Experiment 2 to False Belief (dark grey) and False Photograph (white) stories in the same subject's TPJ-M, independently defined by a greater response to Theory of Mind than to Mechanical Inference stories in Experiment 1, $p<0.0001$ uncorrected. Medium grey indicates fixation. Time-course averaged over four runs.

4. Experiment 2. Average percent signal change in left and right TPJ-M, defined in individual subjects (n=14) as voxels that respond significantly more to False Belief (FB) than to False Photo (FP) stories ($p<0.0001$ uncorrected. Response magnitude for these two conditions is illustrative only, since this data was used to determine the region of interest). In the TPJ-M bilaterally the BOLD response to Physical People stories was significantly lower than to Desire stories ($p<0.05$), and not significantly different from Non-human Description stories ($p>0.1$, repeated measures ANOVA). Response decreases are commonly observed in the TPJ vicinity during demanding non-social tasks (Shulman et al 1997, Gusnard and Raichle 2001).

Figure 1.1.

Figure 1.2

(a)


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
% signal change
Left TPJ-M
Right TPJ-M
ToM
MI
Body
Obj

(b)

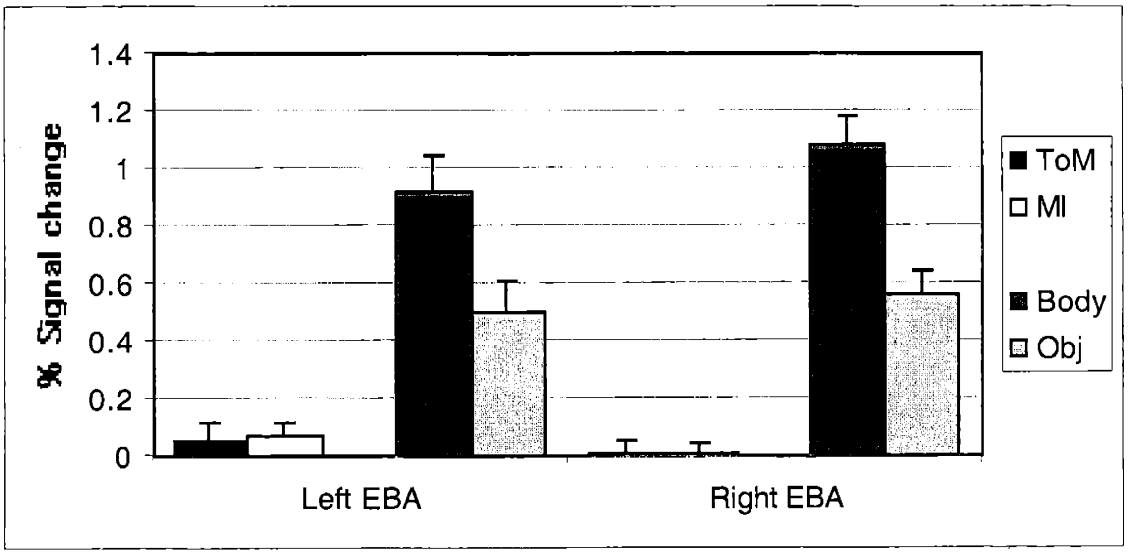

1.4
1.2
1
0.8
0.6
0.4
0.2
0
% Signal change
Left EBA
Right EBA
ToM
MI
Body
Obj

Figure 1.3a.

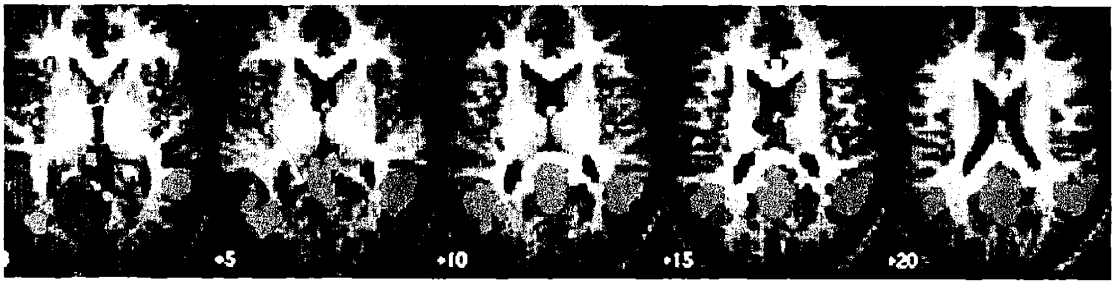


Figure 1.3b.

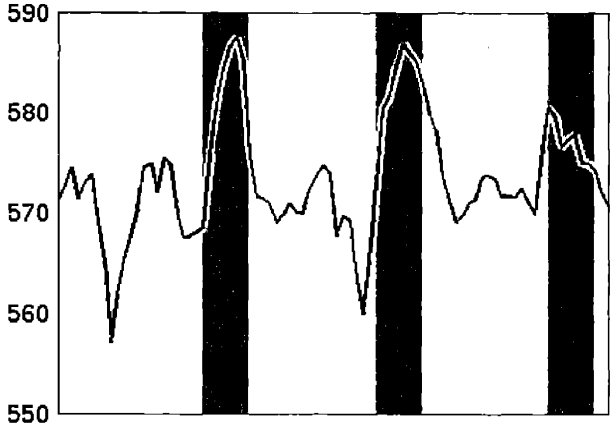
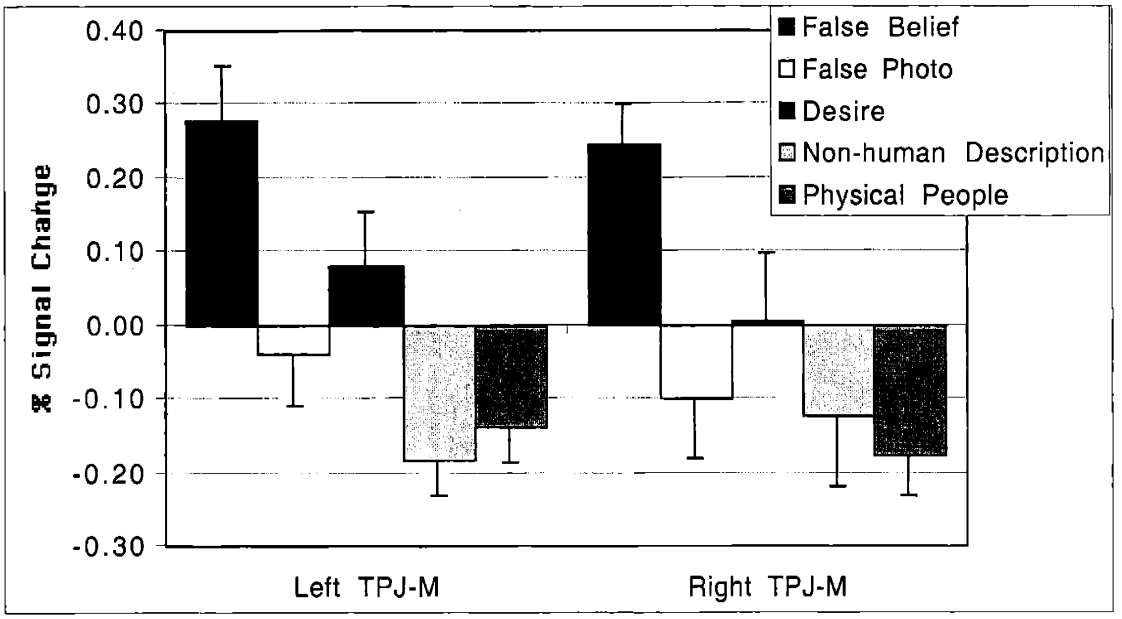


Figure 1.4


0.40
0.30
0.20
0.10
0.00
-0.10
-0.20
-0.30
% Signal Change
False Belief
False Photo
Desire
Non-human Description
Physical People
Left TPJ-M
Right TPJ-M

Appendices

Experiment 1

Instructions: "Read each story silently to your self. Please make sure you understand what is happening; it is more important that you understand the story, than that you go as fast as possible. When you are done reading the story, press the button."

Theory of Mind (ToM) sample story:

A boy is making a paper mache project
for his art class. He spends hours
ripping newspaper into even strips.
Then he goes out to buy flour. His
mother comes home and throws all the
newspaper strips away.

Mechanical Inference (MI) sample story:

A pot of water was left on low heat
yesterday in case anybody wanted tea.
The pot stayed on the heat all night.
Nobody did drink tea, but this morning,
the water was gone.

Human Action sample story:

Jane is walking to work this
morning through a very industrial
area. In one place the crane is
taking up the whole sidewalk. To

get to her building, she has to

take a detour.

Experiment 2.

Instructions: "Please read each story carefully. After each story, you will be given one fill-in-the-blanks question about the story. Underneath will be two words that could fill in the blank. Choose the correct word (to make the sentence true in the story) by pressing the left button to choose the left-hand word, and the right button to choose the right-hand word."

False Belief (FB) sample story:

John told Emily that he had a Porsche.

Actually, his car is a Ford. Emily

doesn't know anything about cars

though, so she believed John.

When Emily sees John's car she

thinks it is a

porsche ford

False Photograph (FP) sample story:

A photograph was taken of an apple hanging

on a tree branch. The film took half an hour to

develop. In the meantime, a strong

wind blew the apple to the ground.

The developed photograph shows the apple on the

ground branch

Desire sample story:

For Susie's birthday, her parents decided

to have a picnic in the park. They wanted

ponies and games on the lawn. If it rained,

the children would have to play inside.

Susie's parents wanted to have her birthday

inside outside.

Physical People sample story:

Emily was always the tallest kid in her

class. In kindergarten she was already

over 4 feet tall. Now that she is in

college she is 6'4". She is a head taller

than the others.

In kindergarten Emily was over

4 ft 6 ft

... tall

Non-human Description sample story:

Nine planets and their moons, plus various
lumps of debris called asteroids and
comets, make up the sun's solar system.
The earth is one of four rocky planets
in the inner solar system.

The solar system has

four nine

... planets.

References

Allison, T., A. Puce, et al. (2000). "Social perception from visual cues: role of the STS region." Trends Cogn Sci **4**(7): 267-278.

Carey S. (1985). Conceptual Change in Childhood. Cambridge, MA: MIT Press

Carruthers P. and Smith P. (1996) Theories of Theories of Mind New York, Cambridge University Press

Bartsch K & Wellman H (1995) Children Talk about the Mind.New York, Oxford University Press, pp 234

Brunet, E., Y. Sarfati, et al. (2000). "A PET investigation of the attribution of intentions with a nonverbal task." Neuroimage **11**(2): 157-66.

Castelli, F., F. Happe, et al. (2000). "Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns." Neuroimage **12**(3): 314-25.

Dennett, D. (1978). "Beliefs about beliefs." Behavioral and Brain Sciences **1**: 568-570.

Dennett, D. (1996) Kinds of Minds: Toward an Understanding of Consciousness New York NY, Basic Books.

Downar, J., A. P. Crawley, et al. (2001). "The effect of task relevance on the cortical response to changes in visual and auditory stimuli: an event-related fMRI study." Neuroimage **14**(6): 1256-67.

Downing, P. E., Y. Jiang, et al. (2001). "A cortical area selective for visual processing of the human body." Science **293**(5539): 2470-3.

Ferstl E. C, von Cramon D. Y. (2002) "What does the frontomedian cortex contribute to language processing: coherence or theory of mind?" Neuroimage. **17**(3):1599-612.

Fletcher, P. C., F. Happe, et al. (1995). "Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension." Cognition **57**(2): 109-28.

Frith, C. D. and U. Frith (1999). "Interacting minds--a biological basis." Science **286**(5445): 1692-5.

Gallagher, H. L. and Frith, C. D. (2003) "Functional imaging of 'theory of mind'" Trends Cog Sci **7**(2): 77-83

Gallagher, H. L., F. Happe, et al. (2000). "Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks." Neuropsychologia **38**(1): 11-21.

Greene, J. and Haidt, J. (2003) "How (and where) does moral judgement work?" Trends Cog Sci **6**(12): 517-23.

Grossman, E., M. Donnelly, et al. (2000). "Brain areas involved in perception of biological motion." J Cogn Neurosci **12**(5): 711-20.

Gusnard, D. A. and M. E. Raichle (2001). "Searching for a baseline: functional imaging and the resting human brain." Nat Rev Neurosci **2**(10): 685-94.

Hoffman, E. A. and J. V. Haxby (2000). "Distinct representations of eye gaze and identity in the distributed human neural system for face perception." Nat Neurosci **3**(1): 80-4.

Leslie, A. (1999). "Theory of Mind" as a Mechanism of Selective Attention. In The New Cognitive Neurosciences. M. Gazzaniga. Cambridge MA, MIT Press.

Malle, B.F., Moses, L.J., & Baldwin, D.A (Eds) (2001) Intentions and intentionality: Foundations of social cognition. Cambridge, MA: MIT Press.

Moll J et al (2002) "The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions." J. Neurosci. **22**: 2730-36

Narumoto, J., T. Okada, et al. (2001). "Attention to emotion modulates fMRI activity in human right superior temporal sulcus." Brain Res Cogn Brain Res **12**(2): 225-31.

Rowe, A. D., P. R. Bullock, et al. (2001). ""Theory of mind" impairments and their relationship to executive functioning following frontal lobe excisions." Brain **124**(Pt 3): 600-16.

Saxe, R., Xiao, D.K. et al (submitted) "Distinct Representations of Bodies, Actions and Thoughts in posterior Superior Temporal Sulcus".

Shulman, G. L., M. Corbetta, et al. (1997). "Top-down modulation of early sensory cortex." Cereb Cortex **7**(3): 193-206.

Stuss, D. T., G. G. Gallup, Jr., et al. (2001). "The frontal lobes are necessary for 'theory of mind'." Brain **124**(Pt 2): 279-86.

Tong, F., Nakayama, N., et al. (2000) "Response Properties of the Human Fusiform Face Area." Cognitive Neuropsychology, **17**: 257-279.

Vogeley, K., P. Bussfeld, et al. (2001). "Mind reading: neural mechanisms of theory of mind and self-perspective." Neuroimage **14**(1 Pt 1): 170-81.

Wellman and Gelman (1992) "Cognitive development: foundational theories of core domains". Annu Rev Psychol. **43**: 337-75.

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

Winston, J. S., B. A. Strange, et al. (2002). "Automatic and intentional brain responses during evaluation of trustworthiness of faces." Nat Neurosci **5**(3): 277-83.

Zaitchik, D. (1990). "When representations conflict with reality: the preschooler's problem with false beliefs and "false" photographs." Cognition **35**(1): 41-68.

Chapter 2. Distinct representations of Bodies, Actions and Thoughts in Posterior Superior Temporal Sulcus

How do we determine why other people do the things they do? Many researchers have proposed a specific role for the posterior superior temporal sulcus (pSTS) in perceiving the behaviour of agents and analysing the goals and outcomes of this behaviour[1,2,3,4]. However, the pSTS region identified in previous studies is large[4,5] and unlikely to be homogenous. Here we test whether the pSTS contains a set of distinct mechanisms for understanding others' actions.

At least three different functions have been attributed to the pSTS. First, previous studies have implicated the pSTS in the visual recognition of human bodies[6] and body movements[2,7]. In the anterior STS of monkeys[8,9,10] and posterior STS of humans[3,5], activity has been reported during observation of hand[7,11], mouth[12], eye[12,13,14,15] and whole body[7,16] movements, displayed as movies[16] or as point-light stimuli[7,17].

Other evidence suggests that the pSTS may integrate the perceived physical structure of a body movement with information about the goals of the action[7,8,18,19,20,21,22]. For instance, heightened pSTS activity is temporally correlated with salient event boundaries (transitions between sub-goals) within an extended action sequence[23]. Also, in the anterior STS of macaque monkeys, a population of neurones responds to the sight of reaching but only when the agent performing the action is seen to be attending to the target position of the reaching[24].

Finally, the pSTS may be involved not merely in recognising *what* an agent is doing, but also in determining *why* she is doing it. A region within the nearby temporo-parietal junction (TPJ-M), responds preferentially to verbal descriptions or cartoons of actions that describe or imply the agent's mental states, i.e. her reasons for acting[25,26,27,28,29].

In short, the pSTS may be involved in three component processes of understanding other people[30]: recognition of a human body and body motion, perceptual analysis of a goal-directed action, and reasoning about a person's mental states. Do these three processes engage the same or distinct regions of human posterior STS? To address this question, we first designed novel stimuli to distinguish the analysis of goal-directed action from the perception of biological motion, based on stimuli previously found to activate neurones in macaque STSa[31,32] and human posterior STS[23]. Subjects watched short movies in which a target person traversed a scene, passing behind a central occluding object (e.g. bookcase). In the critical comparison, we varied how long the person remained invisible (occluded): either the person re-emerged immediately (Short

Occlusion), or else the person remained hidden for a few seconds before re-emerging (Long Occlusion). The visible biological motion was exactly matched (Figure 1). However, the Long Occlusion required subjects to revise their initial action representation (e.g. "the person is walking across the scene") and form a complex representation including two sub-goals ("the person walked to the centre of the scene, and has stopped behind the bookcase"). To foreshadow our results, a region of right posterior STS produced a reliable increased BOLD response to the Long (vs. Short) Occlusion of a walking person. In subsequent manipulations, we further established that this differential activity did not reflect other differences between Long and Short Occlusion movies, such as the extended presence of an occluded person or a break in a smooth motion. The same region was also more strongly engaged by simple geometric animations depicting goal-directed action than by control animations. In keeping with our hypothesis that this region is involved in the representation of observed actions, we refer to this region as the pSTS-VA (Visual analysis of Action).

Second, we tested the functional and anatomical relationship between the pSTS-VA, and two neighbouring regions reported in previous human neuroimaging studies: the Extrastriate Body Area (EBA[6]) which responds to the visual appearance of human bodies, and the TPJ-M[25,26,29] which represents other people's mental states. The pSTS-VA was adjacent to, but distinct from, both the EBA and the TPJ-M. These results suggest that the pSTS and surrounding cortex contain at least three distinct mechanisms for representing others' bodies, actions and thoughts.

Results

A. The pSTS-VA

A single region in right posterior superior temporal sulcus (pSTS-VA) showed a significantly higher BOLD response to a simple walking motion when the target person stopped behind the bookcase (Long Occlusion > Short Occlusion Random Effects n =16, $p<0.0001$ uncorrected, peak voxel: [54 –42 9], $Z = 5.197$, Figure 2a; the differential activity in this region remained significant [$p<0.05$] after a correction for multiple comparisons over the whole brain). Voxels in this region were significantly more active during Long Occlusion compared to Short Occlusion movies in twelve out of sixteen individual subjects ($p<0.0001$ uncorrected in each). There was no significant difference in the response of the pSTS-VA, or of any other region, to a walking person who was briefly occluded (Short Occlusion) versus never occluded (No Occlusion). An analysis of the BOLD time-course revealed no difference in the pSTS-VA response to Long, Short and No Occlusion conditions when the actor was visible; when the actor was invisible the

response was significantly higher during Long Occlusion only (interaction $p<0.001$, repeated measures ANOVA, n=16, Figure 3).

We replicated the basic finding of the pSTS-VA in a second group of subjects with new movies. The BOLD response in the pSTS-VA (an independent group ROI defined by Long > Short Occlusion, Random Effects $p<0.0001$ uncorrected, from the first group of subjects) was higher to Walk Long Occlusion (average percent signal change from fixation (PSC) =0.33), than to Walk Short Occlusion movies (PSC: 0.22, $p<0.005$, paired samples t-test, Figure 3). The existence of the pSTS-VA was confirmed by an independent whole brain analysis of the activity in the second group of subjects (Random Effects n = 14, $p<0.0001$ uncorrected, local maximum: [51 –42 18], $Z=6.01$).

B. Alternative accounts of the pSTS-VA

What does the increased response in the pSTS-VA during Long Occlusion movies reflect? Since the movies in the Long Occlusion and Short Occlusion conditions were visually matched, the differential response of the pSTS-VA to these two conditions cannot be explained by low-level physical differences in the stimuli, nor by a categorical response to biological motion or moving people. We hypothesised that pSTS-VA was involved in the perceptual analysis of action. But there remained at least two alternative features of the Long Occlusion movies that could be responsible for the pSTS-VA response: (1) the extended presence of an occluded person or (2) the break in a smooth motion trajectory. We addressed these possibilities in two subsequent manipulations.

The second group of subjects saw movies in which the intentional appearance of the action was disrupted by having the target person glide passively across the scene, instead of walking. In a behavioural pilot study, naïve observers judged that the walking motion looked more "on purpose" (average score 6.05 of a maximum 7) than the gliding motion (average score 3.3, n=15, $p<0.0001$ paired-samples t-test). Long Occlusion movies with both Walking and Gliding people contained the same break in a smooth motion path, and an occluded person of equal duration. However, the pSTS-VA did not distinguish between Glide Long (PSC: 0.25) and Glide Short Occlusion movies (0.26; $p>0.5$, paired-samples t-test). The interaction in pSTS-VA response between occlusion duration (Long versus Short) and motion type (Walk versus Glide) in a repeated measures ANOVA was significant ($p<0.05$).

One potential weakness of these stimuli was that the passive gliding person looked unnatural and weird (which may explain the unpredicted high response to gliding motion overall in this region). In a third group of subjects, we therefore measured the response of the pSTS-VA to natural passive occlusion events produced by apparent observer motion (i.e. the camera panned across the scene with a central occluding panel

attached). The target person remained completely stationary within his/her environment throughout. As with the gliding people, the pSTS-VA did not respond more to long than short occlusion events caused by observer motion (Observer Motion Long Occlusion PSC: 0.26, $p>0.5$, Observer Motion Short Occlusion PSC: 0.25 paired-samples t-test, n=12). The increased response of the pSTS-VA to long occlusion for walking people was replicated in these subjects in a subsequent localiser scan (Long Occlusion PSC: 0.27; Short Occlusion PSC: 0.22, $p<0.04$, paired-samples t-test, n=11).

Thus in two sets of control conditions, when the overall percept of intentional action was disrupted, the pSTS-VA did not respond to the difference between long and short occlusion movies. We concluded that the pSTS-VA response to long occlusion of a walking person did not reflect the presence of a broken motion trajectory or of an occluded person. Interestingly, the pSTS-VA response to matched short occlusion movies was significantly higher for a walking person (New Walk Short Occlusion PSC: 0.30) than for a stationary person (Observer Motion Short Occlusion: PSC: 0.25, $p<0.005$, paired-samples t-test, n=12, third group of subjects), consistent with a role in the analysis of goal-directed action.

To confirm the association of the pSTS-VA with the perception of goal-directed action, we used a direct comparison of goal-directed versus non-goal-directed motion. We measured the response of the pSTS-VA (independent group ROI from Experiment 1) while subjects watched animations of two-dimensional geometric shapes. In the "goal directed" motion condition, one shape appeared to be an actor moving other inanimate objects to achieve a simple goal (e.g. putting a ball in a box). In the "rotation" condition, a display of the same shapes underwent fast rigid rotation, changing direction of rotation unpredictably. Consistent with our prediction, the pSTS-VA responded robustly to goal-directed motion (PSC: 0.88) and not to rotation (PSC: 0.25, repeated measures t-test $p<0.002$).

C. Dissociating the pSTS-VA, the EBA and the TPJ-M

The pSTS-VA was anatomically and functionally distinct from the nearby right Extrastriate Body Area[6] (EBA). There was no difference in the EBA response to Long Occlusion and Short Occlusion movies (Long Occlusion PSC: 0.39, Short Occlusion PSC: 0.37, $p>0.5$). There was also no significant difference in the right pSTS-VA response to photographs of Bodies (PSC: 0.10) versus non-human Objects (PSC: 0.02, $p>0.1$). In each subject, the EBA was located posterior to the pSTS-VA (mean centre of right EBA in 7 subjects: [52 –66 9], mean centre of pSTS-VA in the same subjects [62 –45 10], Figure 2b).

The pSTS-VA was also doubly dissociated from the TPJ-M ([51 -54 27][29]). In independently defined group ROIs, only the pSTS-VA distinguished between Short and Long Occlusion movies ($p<0.005$). The TPJ-M showed a much larger effect than the pSTS-VA of Belief versus Photo stories. The three-way interaction between region (pSTS-VA versus TPJ-M), task (Stories versus Movies) and effect size (Test versus Control condition) was highly significant (repeated measures ANOVA $p<0.0001$). That is, the difference between Long and Short Occlusion was represented only in the pSTS-VA, and the difference between Belief and Photograph stories was significantly greater in the TPJ-M than in the pSTS-VA (Figure 5).

The pSTS-VA group peak did show a significantly greater response to Belief than Photograph stories (paired samples t-test, $p<0.01$). However, this appeared to be an artefact of a group analysis, in which the variability between subjects led to a blurring of the responses of neighbouring regions. Consistent with this hypothesis, individual subjects analyses revealed that the pSTS-VA response did not differ between Belief (PSC: 0.03) and Photograph stories (PSC: 0.02, paired samples t-test, $p>0.4$, Figure 4). Also in individual subject ROIs, the TPJ-M did not discriminate between Long (PSC: 0.09) and Short Occlusion movies (PSC: 0.16, paired samples t-test, $p>0.1$). The TPJ-M was consistently posterior to the pSTS-VA (Figure 2b).

D. Response to inanimate objects

How would the pSTS-VA respond to the occlusion of inanimate objects? To address this question, the second group of subjects saw inanimate artefacts (e.g. a china vase) that appeared to glide across the room, following the same trajectory as the gliding people; the third group of subjects saw stationary plants undergo passive occlusion from apparent observer motion, matched to the stationary people. There was no consistent effect of occlusion duration for inanimate objects across these two groups in the pSTS-VA. For the gliding objects there was a trend in the direction of a greater response to the Object Long Occlusion (PSC: 0.22) than to the Object Short Occlusion (PSC: 0.16, $p=0.10$, second group of subjects). For the stationary plants, the direction of the effect was reversed (Plant Long Occlusion PSC: 0.17, Plant Short Occlusion PSC: 0.24, $p<0.05$, paired-samples t-test, third group of subjects). These results, along with the results for gliding and stationary people described above, confirm that the pSTS-VA response in this task does not arise from any interruption of a motion path, or from the occlusion of an object.

The pSTS-VA did show an interesting pattern of domain-specificity: a higher response for people than inanimate objects only for moving, not stationary stimuli. In particular, the pSTS-VA responded significantly more to the gliding people than to the

gliding objects (repeated measures ANOVA, main effect of Person > Object, $p<0.05$). On the other hand, there was no main effect of Person>Plant in the Observer Motion condition, replicating the previous finding of no differential response in this region to stationary people compared with stationary objects. A different response profile would be expected in the nearby EBA, and indeed a cluster of voxels in right extrastriate cortex (peak: [57 –60 15]) was significantly more active (Random Effects analysis, n=12, $p<0.0001$ uncorrected) in the observer motion short occlusion condition for a stationary person (PSC: 0.25) than a stationary plant (PSC: 0.12). These conditions did not differ in the pSTS-VA (repeated measures ANOVA interaction of Target [Person>Plant] by Region [EBA versus pSTS-VA] $p<0.006$). Thus, the EBA appears to make a categorical discrimination between people and inanimate objects for both moving and stationary stimuli, while the pSTS-VA does so only for moving stimuli.

Discussion

A. Three distinct regions

The most important finding of this paper is the discovery that (at least) three distinct brain regions near the human right posterior superior temporal sulcus, with different functional profiles, play a role in the perception and understanding of other people. Many papers have reported an association between pSTS activity and representations of human body form and motion, goal-directed action, and other people's mental states[3]. However, previous authors have sought a unifying explanation of all of these patterns and proposed a homogenous role for the pSTS in the understanding of other people[33,34]. Our results suggest a different resolution. Instead of functional homogeneity, we provide evidence for three neighbouring but distinct regions near the right pSTS: the EBA involved in representing human body form, the pSTS-VA involved in processing intentional action, and the TPJ-M which represents the contents of mental states.

The most posterior of the three regions investigated in this paper is the previously-described Extrastriate Body Area (EBA, group peak from ref. 6: [51 –71 1]). The EBA shows a high response to the visual form of a human body, for both moving and stationary stimuli. The EBA also shows some preference for biological over non-biological motion of a human body[6]. However, the EBA did not respond to the presence of an invisible (occluded) person in the Long Occlusion movies, nor to the presence of people in verbal stories, suggesting that the EBA represents the visual appearance of human bodies.

Superior to the EBA is another previously-described region, the TPJ-M (group peak from ref. 29: [51 –54 27]). The TPJ-M responded preferentially when subjects read verbal stories that describe or imply a character's mental states[25,26,29], compared with logically similar stories about non-mental representations[29] (e.g. photographs and maps). The response of the TPJ-M was also significantly higher during stories describing human actions and mental states than during a movies of a walking person, and did not distinguish between different walking actions (e.g. Long versus Short occlusion), suggesting that perception of a goal-directed action was not sufficient for the TPJ-M. Rather the TPJ-M may represent the contents of others' mental states.

Finally, anterior to the TPJ-M we identified a region of human right pSTS that appears to be involved in the perceptual analysis of goal-directed action, referred to here as the pSTS-VA (group peak from the current study: [54 –42 9]). In three groups of subjects, the response of the pSTS-VA to the same visible motion increased significantly, if a person observed walking across a room appeared to stop for a few seconds behind an occluding object (the Long Occlusion condition). The differential response to the Long Occlusion was eliminated when the perception of intentional action was disrupted. We hypothesised that the Long Occlusion required observers to form a more complex action representation than the Short Occlusion; the former requiring at least two distinct sub-goals: walking across the room, and stopping behind the bookcase. A related possibility is that the increased response occurred when subjects revised their initial simple action representation into a new more complex representation due to the unexpected stop. The same region also showed an increased response to the motion of simple geometrical figures when that motion was perceived to constitute a goal-directed action, consistent with a role for the pSTS-VA in the representation of intentional action (cf. ref. 27).

It is striking that although these three regions are adjacent to one another, in individual subject ROIs none of the regions showed any hint of the effect used to define the other two regions. That is, neither the pSTS-VA nor the TPJ-M responded more to stationary human bodies than to familiar objects, the contrast used to define the EBA. Neither the EBA nor the pSTS-VA showed higher activity during reasoning about mental states than during reasoning about non-mental representations (e.g. photographs and maps), the contrast used to define the TPJ-M. Finally, neither the EBA nor the TPJ-M discriminated the Long and Short Occlusion movies used in this study to identify the pSTS-VA. Each of these comparisons was significant in a contrast by region interaction. Thus, though the precise role of these (and other) brain regions in understanding other people may be clarified or modified by future research, our results have clearly established the existence of at least three functionally distinct regions in the vicinity of the posterior superior temporal sulcus.

B. The role of the pSTS-VA

In addition to the triple dissociation of brain regions near the pSTS, in this study we provided an extended characterisation of one of these regions, the pSTS-VA. The profile of the pSTS-VA response was strikingly consonant with the results of single cell recordings in monkey anterior STS[8,10,31,35,36,37]. Both the pSTS-VA and monkey anterior STS showed no difference between walking (articulated) and gliding (rigid) motion of a person, and decreased response during (apparent) observer motion, providing evidence for homology. Also, unlike parietal and premotor neurones[38,39], the STS in both humans and monkeys seems to selectively represent the actions of others, but not of the self[21,40] (but see ref. 41). We proposed a role for the pSTS-VA in the interpretation of goal-directed action.

Biological motion. However, alternative explanations of a greater response to Long compared with Short occlusion movies were possible. A nearby region of the posterior superior temporal sulcus has previously been implicated in the representation of biological motion[16,17,42,43,44]. Did the pSTS-VA response reported here simply reflect the perception of biological motion? There are at least three possible interpretations of the term "biological motion": (1) the characteristic articulated motion of human (or animal) bodies, (2) any motion of a human or animal, or (3) any goal directed action. Only the third version was consistent with the functional profile observed in the pSTS-VA.

The pSTS-VA did not respond preferentially to articulated human body motion. First, the Long and Short Occlusion movies, the contrast used to defined the pSTS-VA above, differed only in the duration of an invisible cessation of motion: the distance and speed of both visible and inferred biological motion were identical in these two conditions. Second, the pSTS-VA did not respond more to biological (articulated walking) than to rigid (gliding) motion of a person. Third, animations of rigid geometric objects, with no articulation, elicited a high response in the pSTS-VA, when these motions constituted goal-directed actions. Thus, while some region of the posterior superior temporal sulcus may have a particular role in the representation of articulated biological motion[16,32,42,43], the pSTS-VA does not.

A different possibility is that the pSTS-VA responds to any motion of a person (or an agent), whether articulated or rigid[42]. This hypothesis could account for the high response to both walking and rigidly gliding humans and to intentionally moving geometrical figures. However, this revised biological motion hypothesis cannot account for the observed asymmetry between walking and gliding: an unexpected stop behind the bookcase produced an increased response in the pSTS-VA only for walking people, when it was perceived to be an intentional action.

Finally the term "biological motion" is used sometimes to refer to any goal-directed action. In this final sense, the biological motion hypothesis does not differ from our own interpretation of the pSTS-VA. Abstract visual descriptions of actions that permit interpretation of intentions appear to be the domain of the pSTS-VA (or STSa in the monkey[24,31,40]), whether or not they are defined by agents moving in an articulated fashion.

Unexpected target detection. A distinct alternative hypothesis is that the pSTS-VA responded simply to the unexpected appearance (or non-reappearance) of the target person from behind the bookcase. Prior reports indicate that the response of a nearby region in the right TPJ increases following detection of simple geometric targets, especially when spatially or temporally unpredicted (e.g. the 'invalid' trials in a standard Posner spatial cueing paradigm[45,46]). The anatomical vicinity of regions involved in social cognition and target detection may reflect a true functional relationship[47], and will be investigated in future work.

However, it is already clear that a simple response to an unexpected target cannot account for all of our results in the pSTS-VA. First, in four sets of control conditions there was no increased response in the pSTS-VA when a target object made an unexpected stop and reappearance. The unexpected stop was significant only when it was perceived as part of an intentional action. Second, the pSTS-VA showed a higher response to a person walking across a room than to a stationary person with a camera pan (observer motion), neither of which involved an unexpected event. Third, the pSTS-VA responded more to gliding people than to gliding objects, when the trajectories of the targets were identical. These latter two patterns are consistent with a role in analysis of human action, but are not predicted by a response to an unexpected target per se.

A revised version of the "unexpected targets" hypothesis might therefore suggest that the pSTS-VA response increased specifically to an unexpected change in a goal-directed action. This revised hypothesis is similar to our own interpretation of the pSTS-VA. A region involved in the representation of intentional action could be recruited for revision of an action representation, when an unexpected change in the observed action occurred. Consistent with this idea, Pelphrey et al[22] reported that the response of a region of right pSTS to perceived changes in a person's gaze direction was enhanced (and extended) when the person looked away from rather than towards a transient target grating. Also, Decety and Chaminade[48] found a higher response in a nearby region when an actor recounted a negative personal experience using positive emotional facial and vocal expression compared with a congruent negative facial expression. In each of these studies, and in our own, reformulation of expectations following an unexpected change in action may have extended activity in a brain region devoted to representing intentional action, the pSTS-VA.

C. Conclusions

The proposal that a region of pSTS is involved in some aspect of representing human bodies, actions or thoughts is relatively uncontroversial. What is striking in our data is the finding that at least three distinct brain regions within that general vicinity are separately responsible for each of these aspects of perceiving other people. In individual subjects, each of the regions (the EBA, pSTS-VA and TPJ-M) were doubly dissociated from the other two, both anatomically and functionally (see also ref. 29).

What exactly is the division of labour between the pSTS-VA and the TPJ-M? We speculate that the pSTS-VA constructs representations of goal-directed actions, whereas the TPJ-M is involved in the abstract analysis of the contents of mental states[30,49,50] (see ref. 51 for evidence from autism and William's syndrome for a distinction between these two levels of analysis). One formulation of this distinction is Leslie's[30] proposed division between the "actional properties" of an agent, those that let it act in pursuit of goals and react to the environment, and the "cognitive properties" of an agent and the contents of her mental states, which include beliefs and desires. Thus the pSTS-VA may reflect a representation of others' "actional properties", which is available to monkeys[8,24,39] and human infants[52,53], while TPJ-M activity may be necessary for a representation of "cognitive properties" only available to (relatively) mature humans[54,55]. This is the first direct evidence we know of in the healthy adult brain of such a two-tiered system for reasoning about other minds[56].

Methods

Fifty healthy right-handed adults (24 women) volunteered or participated in this study for payment: 16 in the first group of subjects, 14 in the second group of subjects, 12 in the third group and 8 in the fourth group. All subjects had normal or corrected to normal vision and gave informed consent to participate in the study.

Subjects were scanned in the Siemens 1.5 T (and 3.0 T – group 4 only) scanner at the MGH-NMR center in Charlestown, MA, using a head coil. Standard echoplanar imaging procedures were used (TR = 2 sec, TE = 30 msec, flip angle 90°). Twenty 5mm thick axial slices covered the whole brain, excluding the cerebellum.

Stimuli consisted of 8.0 second long movies, presented in colour using the Matlab PsychToolbox[57,58] in Quicktime format (QT.mex, 30 frames per second). Each movie began with an empty room, containing a large bookcase. Then, a human actor or target object crossed the room, from one side to the other (left to right in 50% of movies) leaving the room empty once again (Figure 1).

Subjects in group 1 saw 18 movies from each of three conditions. (A fourth condition was not analysed here). **No Occlusion** movies began with 4 seconds of the

empty room. Then an actor emerged, walked across the room for 3.5 (seconds passing in front of the bookcase), leaving the room empty for the final 0.5 seconds of the movie. The long period of empty room was placed at the beginning of the trial (rather than at the end) to ensure that subjects remained equally vigilant throughout the 8.0 seconds of movie in all three conditions. **Short Occlusion** movies used the same timing parameters as No Occlusion movies, except that the actor passed behind the bookcase, rendering him/her very briefly invisible. Finally, in the **Long Occlusion** movies, the room began empty for 1 second, and then the actor appeared and walked to behind the bookcase. Here the actor paused for 3 seconds, and then emerged, and walked off stage. These movies were created by digitally editing the Occlusion movies, so that the observed biological motion in these two conditions was identical.

Movies were presented in blocks of three movies from a single condition, producing blocks of 24.0 seconds. Each scan lasted 464 seconds and contained four blocks of each condition, and five periods of fixation lasting 16.0 seconds. The order of conditions was counterbalanced within and across runs. Half of the movies seen by each subject contained an occlusion, and one quarter of the movies contained a long occlusion. Subjects viewed 2 to 5 scans of these stimuli. Each movie appeared between one and three times over the course of the experiment.

Subjects were instructed to press a button when the actor first appeared on the stage, and when the actor left the stage, in each movie. These responses were monitored during the scanning, to ensure that the subjects were awake and attending. However, due to technical limitations of QT.mex, the timing of the responses could not be recorded for off-line analysis.

Eight subjects (5 women) from group 1 also ran an EBA localiser in the same scan session. Stimuli consisted of twenty grayscale photographs of whole human **bodies** (including faces) in a range of postures, standing and sitting, and twenty photographs of easily recognisable inanimate **objects** (e.g. car, drum, tulip). These photographs were also used in the initial identification of the EBA[6]. (Two other conditions, faces and scrambled objects, were included in the scan but were not analysed here).

Image presentation for the EBA localiser followed the blocked design described in ref. 59 (Experiment 1) except that images were presented at a rate of one every 800 ms (stimulus duration = 500 ms, interstimulus interval = 300 ms), and each scan lasted 336 seconds. Subjects performed a one-back matching task[56]. The right EBA was successfully located in seven of the eight subjects.

The eight remaining subjects from group 1 performed a Stories task in the scanner. Stimuli consisted of twelve stories constructed to fit each of two conditions:

Belief stories required subjects to judge the content of a mental representation, while **Photograph** stories required subjects to judge the content of a physical representation, such as a photograph or a map. After each story a two-alternative forced choice 'fill-in-the-blank' question was presented for 4 s. The question consisted of a single sentence with a word missing, presented above two alternative completions on the left and right side of the screen. Subjects pressed the left-hand response button if the word on the left completed the sentence to fit the story, and the right-hand button to choose the word on the right. Each story was presented to each subject only once. Subjects were given three practice trials before going into the scanner: two **Belief** trials, and one **Photograph** trial. Stimulus presentation for the stories followed the blocked design described in ref. 29 (Experiment 2a). The right TPJ-M activation from the **Belief** stories was successfully identified in seven of the eight subjects.

Subjects in group 2 saw 6 movies in each of eight conditions: **No Occlusion** movies were constructed as in the No Occlusion condition above. Two sets of No Occlusion movies were included in the scan: the second set of No Occlusion movies, **No Occlusion2,** was included to decrease the overall proportion of movies containing an occlusion event. No Occlusion2 movies were not exactly matched to the others in this experiment, and so the response to these stimuli was not analysed. **Walk Long Occlusion** and **Walk Short Occlusion** were designed to match the Long Occlusion and Short Occlusion conditions of Experiment 1, respectively. (These movies used slightly lower velocities and different actors than original movies, to test the generality of brain response beyond a particular set of stimuli.) In the **Glide Short Occlusion** movies, a still cut-out of the actor in profile, facing the direction of motion, moved smoothly across the room with the same velocity and timing as the walking actor in the **Walk Short Occlusion** movies, and undergoing apparent occlusion (gradual occlusion and accretion) at the bookcase edges. **Glide Long Occlusion** movies were created to match the timing parameters of the **Walk Long Occlusion** condition, with the same gliding motion. Finally **Object Short Occlusion** and **Object Long Occlusion** movies were created, as in the **Glide** movies, but replacing the cut-out of a person with a non-human object of similar aspect ratio to the human targets.

Each movie appeared once in each run; each subject saw two runs of these stimuli. The blocked design used the same parameters as the first group of subjects. Subjects were instructed to press a button when the target actor/object first appeared on the stage, and when the target reappeared from occlusion. This task was designed to demand similar spatial attention to the empty room, at the beginning of the Short occlusion movies, and in the middle of the Long occlusion movies, since in each case the subject would be monitoring in order to make a response.

Subjects in group 3 saw 8 movies in each of five conditions. In these movies, the scenes themselves did not include an occluding object. Instead, throughout all of the stimuli a black panel (1/4 the width of the movie) blocked the subject's view of the centre of the scene, like a frame running down the centre of a train window. In **Observer Motion Short Occlusion** and **Observer Motion Long Occlusion** movies, the target was a person standing still, facing the camera. At the start of the film, the camera pointed off to one side (50% left, 50% right), showing an empty scene. Then, the camera panned past the target person so that the person moved smoothly across the screen but remained visibly stationary with respect to his/her environment. The motion of the camera paused at the centre of the scene, occluding the target, for 0.5 seconds (Short Occlusion) or 3 seconds (Long Occlusion). **Plant Short Occlusion** and **Plant Long Occlusion** were exactly matched to Observer Motion Short and Long movies, except that the person was replaced by a potted plant with approximately the same aspect ratio. Plants were chosen because they are familiar and definitely inanimate. We also included matched **New Walk Short Occlusion** movies, in order to establish the magnitude of response to intentional action in these subjects.

Movies were presented in blocks of three movies from a single condition, producing blocks of 24.0 seconds. The run lasted 368 seconds and contained two blocks of each condition, interleaved with five periods of fixation lasting 16.0 seconds. The order of conditions was counterbalanced within and across runs. Subjects were instructed as in group 2, except that a brief asterisk at the start of each movie cued the subject to the side of appearance of the target, to further reduce confounds with spatial attention.

Then, in two separate "replication" runs subjects in group 3 saw the **Long Occlusion** and **Short Occlusion** movies from group 1. These movies were presented in alternating blocks of three movies from a single condition, interleaved with a fixation of 16.0 seconds. Each run lasted 228 seconds. Eleven subjects saw two runs of these stimuli; for one subject, the scan session had to be terminated before these runs.

In order to allow a direct comparison across tasks and regions, eight subjects from group 3 also ran the Stories task. The data from the "replication" movies scan, and from the stories scan, were analysed in independent ROIs, each defined as a sphere of radius 3mm around the peak activation from a separate group of subjects. The ROI from the movies task was from group 1 (above, [54 –42 9], n= 16, Random Effects $p<0.05$ corrected). The ROI from the stories task was defined around the peak from the False Belief > False Photograph comparison in ref. 29 (Experiment 2, [54 –51 18], n= 21, Random Effects $p<0.05$ corrected).

The eight subjects in the fourth group saw **Goal-Directed** and **Rotation** motion in simple geometrical animations. On a debriefing questionnaire, all subjects described the goal-directed motion as depicting "actions," "interaction" or "problem solving", and the rotation as "moving in circles" or "rotating". Each animation lasted 6 seconds, presented in blocks of two animations from the same condition. Timing parameters and design were the same as in the Stories task. Each subject saw two scans (a total of 12 animations in each condition). Subjects were instructed to watch the animations carefully in order to maximise their performance on a memory test administered immediately after the scan. Seven subjects performed perfectly on the subsequent recognition memory task; the eighth subject wrongly identified one new rotation animation as old. All subjects reported that the memory task was harder for the rotating shapes. These subjects also saw the same movies as the subjects in group 1, in two separate scans.

MRI data were analysed using SPM 99 and in-house software.

Figure Legends

1. **Schematic illustration of stimuli from the two critical conditions in Experiment 1, 'Short Occlusion', and 'Long Occlusion'.** In each movie, a person walked across the scene, passing behind a large bookcase. Movies in the 'Short Occlusion' and 'Long Occlusion' conditions were constructed from identical movie fragments. The five components of each movie are illustrated in (**a**): (1) an empty room before the actor emerged, (2) the actor walked to the bookcase, (3) an empty room while the actor was occluded, (4) the actor walked from the bookcase to offstage, and (5) the empty room again. 'Long Occlusion' and 'Short Occlusion' movies differed only in the relative duration of the empty scene in (1) and (3), as illustrated in the timeline of a single trial from each condition (**b**). In Short Occlusion movies, the room was empty only briefly while the actor passed behind the bookcase. In the Long Occlusion movies, the occlusion was extended in time, and the duration of the empty room at the start of the trial was brief. Each trial lasted 8 seconds.

2. **The PSTS-VA, TPJ-M and EBA**. (**a**) The region of superior temporal cortex that was more active during 'Long Occlusion' than during 'Short Occlusion' movies in group 1 (pSTS-VA [Visual analysis of Action]). Shown here are all voxels passing a threshold of $p<0.0001$ uncorrected in a Random Effects analysis of 16 subjects. (**b**) Regions of interest for the EBA (red), TPJ-M (blue), and pSTS-VA (white) in one subject (all $p<0.0001$ uncorrected). In this subject, a few voxels overlapped between the TPJ-M and pSTS-VA (shown in green).

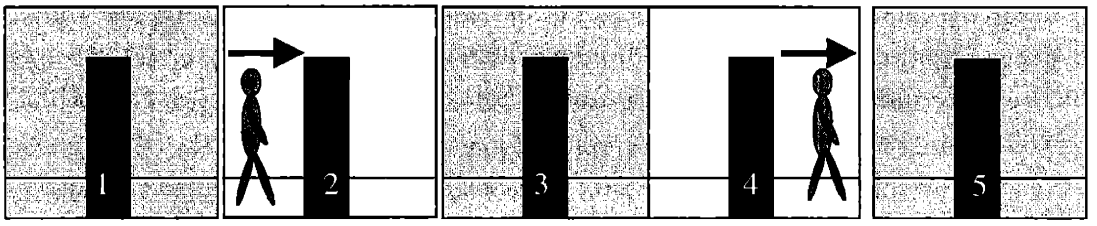
3. **Visible versus Invisible.** Response of the pSTS-VA (group ROI) for phases of each trial when the actor was visible, or invisible but anticipated (group 1, n=16). The response of the pSTS-VA to the visible biological motion in the No Occlusion, Short Occlusion and Long Occlusion movies (phase 2 in Figure 1) did not differ. However, when the person was invisible, the pSTS-VA responded significantly more to the occluded person in the Long Occlusion movies (phase 3), than to the anticipated but not yet present person (phase 1) in the No Occlusion and Short Occlusion movies (both paired-samples t-test $p<0.001$; interaction of Visibility by Occlusion Duration $p<0.002$, repeated measures ANOVA). Note that the visual displays were identical across conditions, for both visible and invisible phases.

4. **Walk versus Glide.** Percent signal change from fixation of the pSTS-VA (independent group ROI from group 1), to five new conditions in group 2 (n = 14). Dark red bars indicate a long occlusion; light grey bars indicate an uninterrupted walking path (short or no occlusion). Two different motions were observed: Walking and Gliding people (static cut-out human bodies moving passively across the room).

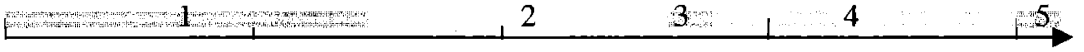
The PSTS-VA response was significantly higher for the Long occlusion only for Walking people ($p < 0.005$, interaction between motion (Walk versus Glide) and occlusion duration (Long versus Short) $p<0.05$). Bars indicate standard error.

5. **Dissociating the pSTS-VA and the TPJ-M.** Response (percent signal change from fixation) of the pSTS-VA and TPJ-M to movie (red and grey bars) and story (blue bars) stimuli (n=8). A sphere of 3mm was defined around the peak voxel in each task ($p<0.05$ corrected), from a different set of subjects: the PSTS-VA peak [54 –42 9] was defined by group 1 of the current study, and the TPJ-M peak [51 –54 18] was defined by subjects in a separate study (Saxe and Kanwisher, submitted). Only the pSTS-VA distinguished between Short and Long Occlusion movies ($p<0.005$). The TPJ region showed a much larger effect of Belief versus Photo stories than the pSTS-VA; the three way interaction between region, task and effect size was highly significant (repeated measures ANOVA, $p<0.0001$). Striped blue bars show the response of individually tailored ROIs in a separate group of subjects (n=7). When anatomical variability was neutralised, the pSTS-VA region did not discriminate between Belief and Photograph stories ($p>0.3$). Lines indicate standard error.

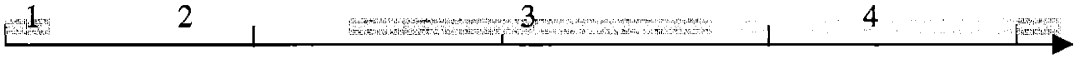
Figure 1 (a)



Short Occlusion:



Long Occlusion:



Timeline of a single trial

Figure 2

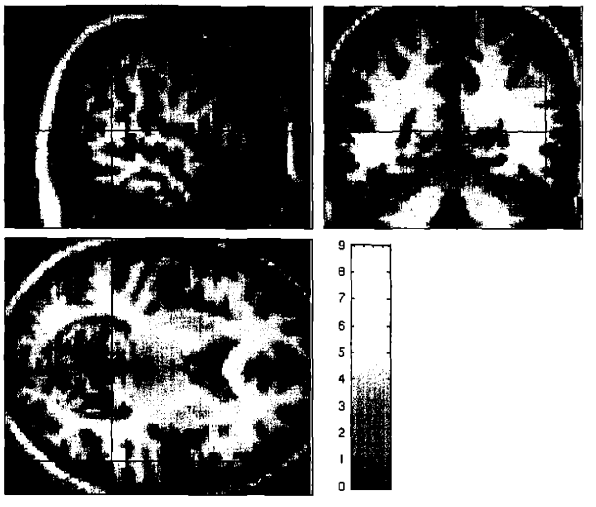


Figure 3

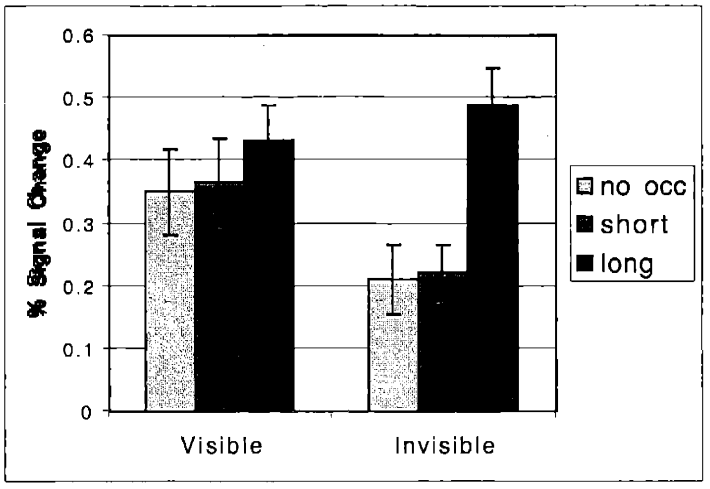

0.6
0.5
0.4
0.3
0.2
0.1
0
% Signal Change
Visible
Invisible
no occ
short
long

Figure 4

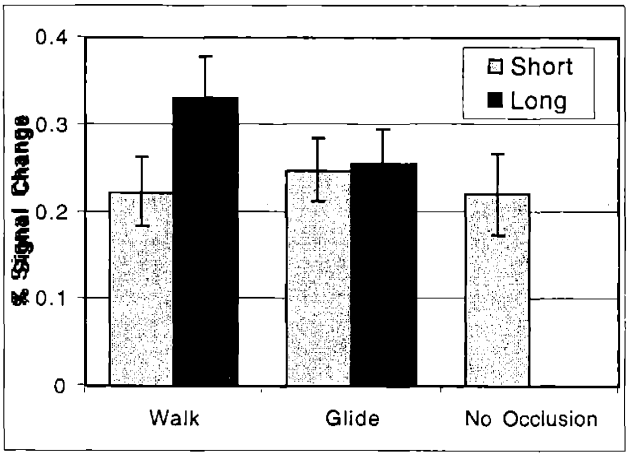
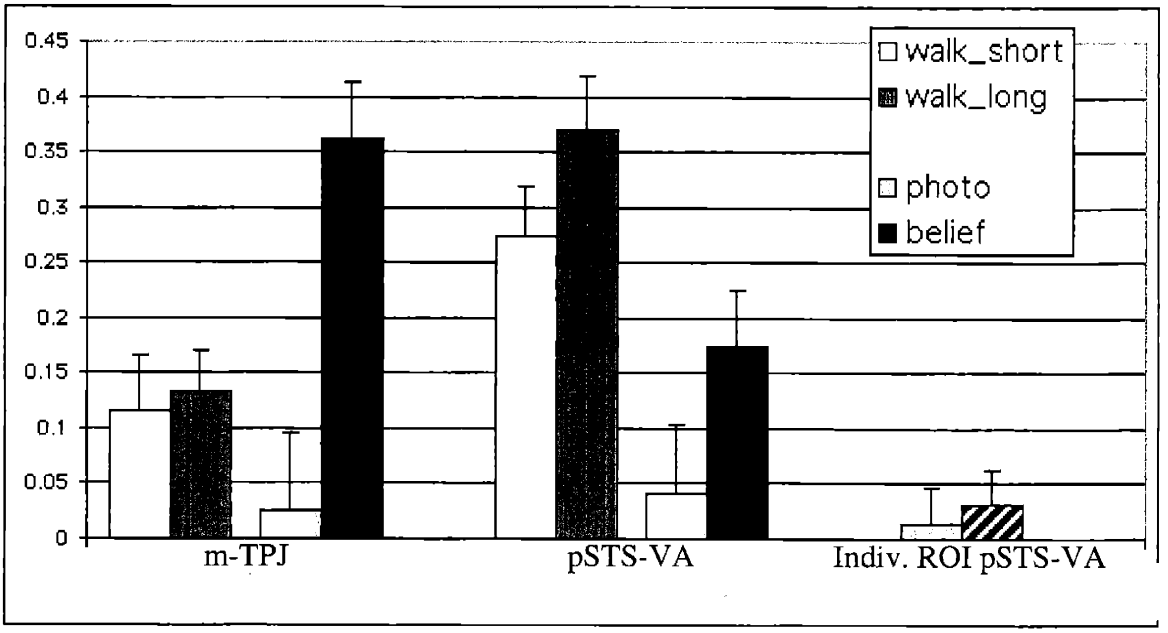

0.4
0.3
0.2
0.1
0
% Signal Change
Short
Long
Walk
Glide
No Occlusion

Figure 5


0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
walk_short
walk_long
photo
belief
m-TPJ
pSTS-VA
Indiv. ROI pSTS-VA

References

1. Brothers, L. (1997). Friday's Footprint: How society shapes the human mind. New York, Oxford University Press.

2. Frith, C. D. and U. Frith (1999). "Interacting minds--a biological basis." Science **286**(5445): 1692-5.

3. Allison, T., A. Puce, et al. (2000). "Social perception from visual cues: role of the STS region." Trends Cogn Sci **4**(7): 267-278.

4. Haxby, J. V., E. A. Hoffman, et al. (2002). "Human neural systems for face recognition and social communication." Biol Psychiatry **51**(1): 59-67.

5. Grezes, J. and J. Decety (2001). "Functional anatomy of execution, mental simulation, observation, and verb generation of actions: a meta-analysis." Hum Brain Mapp **12**(1): 1-19.

6. Downing, P. E., Y. Jiang, et al. (2001). "A cortical area selective for visual processing of the human body." Science **293**(5539): 2470-3.

7. Bonda, E., M. Petrides, et al. (1996). "Specific involvement of human parietal systems and the amygdala in the perception of biological motion." J Neurosci **16**(11): 3737-44.

8. Perrett, D. I., Harries, M., Chitty, A.J. and Mistlin, A.J. (1990). Three stages in the classification of body movements by visual neurones. *Images and Understanding*. H.B. Barlow ,C. Blakemore. and M. Weston.-Smith,. (Eds) Cambridge, Cambridge University Press**:** 94-108.

9. Oram, M. W. and D. I. Perrett (1996). "Integration of form and motion in the anterior superior temporal polysensory area (STPa) of the macaque monkey." J Neurophysiol **76**(1): 109-29.

10. Jellema, T. and D. I. Perrett (2002). Coding of visible and hidden actions. Common mechanisms in Perception and Action. W. Prinz and B. Hommel. Oxford, Oxford University Press. **XIX:** 356-380.

11. Grezes, J. and J. Decety (2002). "Does visual perception of object afford action? Evidence from a neuroimaging study." Neuropsychologia **40**(2): 212-22.

12. Puce, A., T. Allison, et al. (1998). "Temporal cortex activation in humans viewing eye and mouth movements." J Neurosci **18**(6): 2188-99.

13. Watanabe, S., R. Kakigi, et al. (2001). "Occipitotemporal activity elicited by viewing eye movements: a magnetoencephalographic study." Neuroimage **13**(2): 351-63.

14. Hoffman, E. A. and J. V. Haxby (2000). "Distinct representations of eye gaze and identity in the distributed human neural system for face perception." Nat Neurosci **3**(1): 80-4.

15. Wicker, B., F. Michel, et al. (1998). "Brain regions involved in the perception of gaze: a PET study." Neuroimage **8**(2): 221-7.

16. Grossman, E. D. and R. Blake (2002). "Brain Areas Active during Visual Perception of Biological Motion." Neuron **35**(6): 1167-75.

17. Grossman, E. D. and R. Blake (2001). "Brain activity evoked by inverted and imagined biological motion." Vision Res **41**(10-11): 1475-82.

18. Decety, J., J. Grezes, et al. (1997). "Brain activity during observation of actions. Influence of action content and subject's strategy." Brain **120**(Pt 10): 1763-77.

19. Grezes, J., N. Costes, et al. (1998). "Top down effect of the strategy to imitate on the brain areas engaged in perception of biological motion: a PET investigation." Cogn Neuropsychol **15**: 553-582.

20. Perani, D., F. Fazio, et al. (2001). "Different brain correlates for watching real and virtual hand actions." Neuroimage **14**(3): 749-58.

21. Decety, J., T. Chaminade, et al. (2002). "A PET exploration of the neural mechanisms involved in reciprocal imitation." Neuroimage **15**(1): 265-72.

22. Pelphrey, K. A., J. D. Singerman, et al. (2003). "Brain activation evoked by perception of gaze shifts: the influence of context." Neuropsychologia **41**(2): 156-70.

23. Zacks, J. M., T. S. Braver, et al. (2001). "Human brain activity time-locked to perceptual event boundaries." Nat Neurosci **4**(6): 651-5.

24. Jellema, T., C. I. Baker, et al. (2000). "Neural representation for the perception of the intentionality of actions." Brain Cogn **44**(2): 280-302.

25. Fletcher, P. C., F. Happe, et al. (1995). "Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension." Cognition **57**(2): 109-28.

26. Gallagher, H. L., F. Happe, et al. (2000). "Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks." Neuropsychologia **38**(1): 11-21.

27. Castelli, F., F. Happe, et al. (2000). "Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns." Neuroimage **12**(3): 314-25.

28. Vogeley, K., P. Bussfeld, et al. (2001). "Mind reading: neural mechanisms of theory of mind and self-perspective." Neuroimage **14**(1 Pt 1): 170-81.

29. Saxe, R. and N. Kanwisher (In Press). "People thinking about thinking people: the role of the temporo-parietal junction in Theory of Mind." Neuroimage.

30. Leslie, A. (1994). A theory of ToMM, ToBy, and Agency: Core architecture and domain specificity. Mapping the Mind: Domain specificity in cognition and culture. L. Hirschfeld and S. Gelman. New York, Cambridge University Press: 119-148.

31. Baker, C. I., C. Keysers, et al. (2001). "Neuronal representation of disappearing and hidden objects in temporal cortex of the macaque." Exp Brain Res **140**(3): 375-81.

32. Oram, M. and D. Perrett (1994). "Responses of anterior superior temporal polysensory area (STPa) neurones to "biological motion" stimuli." J. Cognitive Neuroscience **6**: 99-116.

33. Frith U, Frith CD. 2003. "Development and neurophysiology of mentalizing". Philos. Trans. R. Soc. Lond. B. Biol. Sci. **358**: 459-73

34. Gallagher HL, Frith CD. 2003. "Functional imaging of 'theory of mind'". Trends Cogn. Sci. **7**: 77-83

35. Hietanen, J. K. and D. I. Perrett (1996). "Motion sensitive cells in the macaque superior temporal polysensory area: response discrimination between self-generated and externally generated pattern motion." Behav Brain Res **76**(1-2): 155-67.

36. Hietanen, J. K. and D. I. Perrett (1996). "A comparison of visual responses to object- and ego-motion in the macaque superior temporal polysensory area." Exp Brain Res **108**(2): 341-5.

37. Jellema, T. and D. I. Perrett (2003). "Perceptual history influences neural responses to face and body postures." J. Cognitive Neuroscience **(Accepted pending revision)**.

38. Kohler, E., C. Keysers, et al. (2002). "Hearing sounds, understanding actions: action representation in mirror neurons." Science **297**(5582): 846-8.

39. Gallese, V., L. Fadiga, et al. (2002). Action representation and the inferior parietal lobule. Common mechanisms in Perception and Action. W. Prinz and B. Hommel. Oxford, Oxford University Press. **XIX:** 247-266.

40. Hietanen, J. K. and D. I. Perrett (1993). "Motion sensitive cells in the macaque superior temporal polysensory area. I. Lack of response to the sight of the animal's own limb movement." Exp Brain Res **93**(1): 117-28.

41. Iacoboni, M., L. M. Koski, et al. (2001). "Reafferent copies of imitated actions in the right superior temporal cortex." Proc Natl Acad Sci U S A **98**(24): 13995-9.

42. Beauchamp M.S., K.E. Lee, J.V. Haxby, A. Martin (2002) "Parallel visual motion processing streams for manipulable objects and human movements." Neuron. **34**(1):149-59.

43. Grezes J., P. Fonlupt, B. Bertenthal, C. Delon-Martin, C. Segebarth, J. Decety (2001) "Does perception of biological motion rely on specific brain regions?" Neuroimage. **13**(5):775-85.

44. Vaina L.M., J. Solomon, S. Chowdhury, P. Sinha, J.W. Belliveau (2001) "Functional neuroanatomy of biological motion perception in humans." Proc Natl Acad Sci U S A. **98**(20):11656-61.

45. Corbetta, M., J. M. Kincade, et al. (2000). "Voluntary orienting is dissociated from target detection in human posterior parietal cortex." Nat Neurosci **3**(3): 292-7.

46. Posner, M. I., J. A. Walker, et al. (1984). "Effects of parietal injury on covert orienting of attention." J Neurosci **4**(7): 1863-74.

47. Leslie, A. (2000). 'Theory of Mind' as a mechanism of selective attention. The New Cognitive Neurosciences. M. Gazzaniga. Cambridge, MA, MIT Press**:** 1235-1247.

48. Decety J, T. Chaminade (2003) "Neural correlates of feeling sympathy". Neuropsychologia **41**: 127-38

49. Wellman, H. M. (1990). The Child's Theory of Mind. Cambridge MA, MIT Press.

50. Baron-Cohen, S. (1994). "How to build a baby that can read minds: Cognitive mechanisms in mindreading." Cahiers de Psychologie **13**: 513-552.

51. Tager-Flusberg, H. and K. Sullivan (2000). "A componential view of theory of mind: evidence from Williams syndrome." Cognition **76**(1): 59-90.

52. Baldwin D.A., J.A. Baird (2001) "Discerning intentions in dynamic human action". Trends Cogn. Sci. **5**: 171-78

53. Woodward AL. (1998) "Infants selectively encode the goal object of an actor's reach". Cognition **69**:1-34

54. Flavell JH. (1999) "Cognitive development: children's knowledge about the mind". Annu. Rev. Psychol. **50**: 21-45

55. Wellman H.M., D. Cross , and J. Watson (2001). "Meta-analysis of theory-of-mind development: the truth about false belief". Child Dev. **72**: 655-84.

56. Saxe, R., S. Carey and N. Kanwisher (In Press) "Understanding other minds: linking developmental psychology and functional neuroimaging." Annu. Rev. Psychol

57. Brainard, D. H. (1997). "The Psychophysics Toolbox." Spat Vis **10**(4): 433-6.

58. Pelli, D. G. (1997). "The VideoToolbox software for visual psychophysics: transforming numbers into movies." Spat Vis **10**(4): 437-42.

59. Tong, F., N. Nakayama, et al. (2000). "Response properties of the Human Fusiform Face Area." Cognitive Neuropsychology **17**: 257-279.

Chapter 3. Understanding other minds: linking developmental psychology and functional neuroimaging.

1 Introduction

Unlike behaviorists, normal adults attribute to one another (and to themselves) unobservable internal mental states, like goals, thoughts, and feelings, and use these to explain and predict behaviour. This human capacity for reasoning about the mental causes of action is called a "theory of mind". In the last twenty-five years, theory of mind has become a major topic of research, initially in developmental psychology and subsequently in other fields including social psychology, philosophy and ethology.

Most recently, a new method has joined the pack: functional brain imaging (especially functional magnetic resonance imaging [fMRI]). The BOLD (Blood Oxygenation Level Dependent) signal measured by fRMI gives scientists unprecedented access to the hemodynamic changes (and indirectly to the neural activity) in the brain that are associated with psychological processes. FMRI may have by now exceeded all other techniques in psychology in terms of expense, growth rate and public visibility. But many have wondered how much this new technology has actually contributed to the study of human cognition, and when if ever a finding from functional neuroimaging has constrained a cognitive theory.

In this chapter, I will take the human theory of mind as a case study of real theoretical exchange between studies (and scientists) using functional neuroimaging, and those using the more established techniques of developmental psychology. As such, the literature reviewed here will necessarily be selective. For more complete coverage of related research in these fields, I refer the reader to the many excellent recent reviews (cognitive neuroscience: Decety and Grezes 1999, Allison, Puce and McCarthy 2000, Adolphs 2001, Blakemore and Decety 2001, Frith 2001, Adolphs 2002, Siegal and Varley 2002, Adolphs 2003, Frith and Frith 2003, Gallagher and Frith 2003, Gallese 2003, Puce and Perrett 2003; developmental psychology: Flavell 1999, Johnson 2000, Wellman and Lagattuta 2000, Wellman, Cross and Watson 2001, Baldwin and Baird 2001, Bartsch 2002, Csibra 2003, Meltzoff and Decety 2003, Johnson 2003).

This review is divided in two main sections, following the substantial evidence for at least two distinct stages in the development of theory of mind (see Figure 3.1a). The first half of the review deals with belief attribution. Around age 3 or 4 children begin to attribute representational epistemic mental states –thoughts, beliefs, and knowledge – to themselves and others. The second half of the review considers the earlier-developing mentalistic reasoning that occurs before age 3. Toddlers do reason about the mind and

human behavior, but they do so with a more limited repertoire of mental state concepts including desires, perceptions and emotions. Within each of these two sections of the review, I will first set out a summary of theoretical questions and findings emerging from developmental psychology, and then consider the contributions to answering these questions that has been or could be made by functional neuroimaging.

Functional neuroimaging is particularly well suited to resolve questions of whether two tasks or processes engage common or distinct mechanisms. For instance, it has been suggested that the development of a concept of belief depends critically on the ability to represent sentence-complement syntax (e.g. de Villiers 2000). However, this dependence could arise either because children cannot learn about beliefs until they can understand the language adults use to talk about the mind, or it could arise because belief attribution is truly dependent on the cognitive and neural mechanisms for parsing sentence complement syntax (even in adulthood). Functional neuroimaging allows us to ask whether these two tasks recruit the same or different regions of the brain[3]. If different brain regions are involved in belief attribution and syntax, then it is less likely that a single functional mechanism is responsible for both. While this use of neuroimaging is potentially powerful in answering fundamental questions about cognition, it is subject to several pitfalls and ambiguities, which I discuss briefly in the next section, before turning to the substance of the review.

2. Functional neuroimaging: standards of evidence and inference.

If we are to take a finding that two different tasks activate the same brain region as evidence that common psychological mechanisms are engaged in the two tasks, then we must consider two questions. First, what counts as the same brain region, and what kind of data can support a claim of common or distinct activations? Second what is the relationship between brain regions and psychological mechanisms?

The location of an activation in the brain is often specified by general region (e.g., the "occipital pole", or the "temporoparietal junction"), or by the gyrus or sulcus where

---

[3] All of the functional imaging discussed in this paper used human adult subjects. In some cases, the adult results alone are sufficient to constrain theories, as this review illustrates. However there are other questions and hypotheses that could only be tested in the brains of infants and children themselves. For instance, are the brain regions involved in response conflict and in belief attribution independent in young children just beginning to attribute beliefs, as they are in adults? fMRI in children and infants is possible (e.g. Born et al 1998, Casey, Thomas and McCandliss 2001, Burgund et al 2002, Dehaene-Lambertz, Dehaene and Hertz-Pannier 2002), but is not widespread. In particular, there are no published fMRI studies yet of children between 10 and 48 months old, the critical age range for theory of mind development, and no fMRI studies of theory of mind in any paediatric population.

the activation is found (e.g. the fusiform gyrus, or the intraparietal sulcus). These descriptors can be useful, but are not very precise, as each one spans ten or more square centimeters of cortex. Such large regions are likely to encompass many functionally distinct areas, as seen for example in extrastriate cortex where areas such as the visual motion area MT (Tootell et. al, 1995) or the fusiform face area (Kanwisher, Mcdermott, & Chun, 1997; McCarthy et. al, 1997) are typically one or two square centimeters in size. Thus even if two tasks produce activations within the same general region of the brain, their activations may not overlap at all. (Imagine concluding that the brain does not contain distinct motor representations of hands and feet, because both hand and foot movements are coordinated by "primary motor cortex".) A further problem is that because individual brains differ from each other physically, there is no theory-neutral way to precisely specify what counts as the "same place" in two different brains.

These problems can be avoided in analyses of individual subjects. The strongest evidence for engagement of the same brain region by two different tasks arises when the very same voxels in the same subject's brain (preferably from the same scanning session) are shown to be significantly activated by each of two different tasks. While some neuroimaging studies meet this high standard, many do not.

A related approach is to first functionally define a region of interest (ROI) individually within each subject based on a particular task comparison (or "localizer" scan), then pool over the voxels in that ROI to ask whether a second task activates the same ROI. This approach avoids the problem of having to register different brains, while making possible statistical analyses over multiple subjects' ROIs. However, in choosing a localizer to define an ROI the researcher is making an ontological assumption that this localizer contrast picks out a meaningful functional unit in the brain (i.e., a "natural kind"). Like other ontological assumptions in science, the utility of a particular functionally-defined ROI is determined by the consistency of the data that emerge from it and the richness of the theoretical progress those data support.

Most common in the neuroimaging literature are group analyses, in which brain images from a dozen or more individuals are aligned (as best as possible) into a common space, and statistical analyses are then conducted across subjects. This method enables one to test whether an activation pattern is consistent across subjects, and it can sometimes provide greater statistical power than individual-subject analyses. However, it comes at the cost of blurring of activation maps due to the necessarily imperfect registration across physically different brains. Individuals vary not only in their physical anatomy but also in their functional anatomy, producing yet more blurring in group-averaged data. Thus, activations that may be completely nonoverlapping within each individual could be highly overlapping when the same data are averaged across subjects.

This problem is exacerbated when comparing activations across subject groups or across studies.

Thus, although group-analyzed data can be informative, any claim that two tasks activate the same place in the brain are most convincing when they are based on analyses within individual subjects or within individually-defined ROIs. The least convincing evidence for common mechanisms comes when each of two tasks produces an activation somewhere within the same large anatomical region (e.g., the Temporo-Parietal Junction or Superior Temporal Sulcus).

The opposite inference, that distinct cortical regions are engaged by two different tasks, is subject to a different and surprisingly common error. Often researchers argue that two tasks engage distinct mechanisms because Task A activates Region X significantly (compared to some control condition) whereas Task B does not. Such arguments are not valid on their own: a difference in significances is not a significant difference. To argue for differential activation of Region X by Task A and Task B, it is necessary to directly compare the activation for the two tasks.

Even with the proper statistical evidence, difficult theoretical issues remain. Psychological theories concern psychological processes whereas neuroimaging data can only test the activation of cortical voxels. What is the relationship between the two? A typical voxel in a neuroimaging study contains hundreds of thousands of neurons, so the common activation of a voxel by two different tasks could arise even if completely distinct neural populations within that voxel are engaged by the two tasks. Thus caution is required when we infer common mechanisms from common activations. That said, if a whole cluster of adjacent voxels shows activation for each of two different tasks but not for many others, it is a reasonable guess that even if distinct neural populations are involved, they are likely to be functionally related.

Caution is also required when making the opposite kind of inference. If each of two tasks activates a distinct and nonoverlapping cortical region, it is clear that different physical brain hardware is engaged by the two tasks, but this need not imply that qualitatively different psychological processes are involved in the two tasks. Consider a stimulus falling in the upper visual field versus the lower visual field: distinct and nonoverlapping regions within primary visual cortex would be activated, but the kind of processing that goes on in each is presumably very similar. Deciding whether dissociations at the cortical level correspond to qualitative differences in processing at the psychological level is a difficult judgement call that requires consideration of the wider theoretical context (e.g., does it make sense computationally that different processes may

be engaged by these two tasks?) and the available empirical evidence (e.g., what other tasks activate these two regions?).

Armed with these cautionary notes about the use of neuroimaging data to resolve psychological questions, I turn now to a review of the literature on the development of theory of mind.

3.1 Developmental Psychology: Belief Attribution

An understanding of other people's thoughts and beliefs plays a central role in adult reasoning about the causes of other people's behaviour (e.g. "she spilled the coffee because she thought the cup was empty", "he's going home because he thinks he left his keys there", Malle 2001) but not, of course, in reasoning about the causes of mechanical events like clockwork or the tides. Given the centrality of beliefs in adult explanations of actions, we might expect that as soon as children begin to explain their own and other's actions, they would use this powerful notion of belief to do so. On the contrary, it is now well established that children do not begin to use beliefs to explain actions until relatively late in development, at 3 or 4 years old.

The most common test of children's ability to explain an action with reference to the actor's belief is the "False Belief" task (Wimmer and Perner 1983; for reviews of this literature see Flavell 1999, Wellman and Lagattuta 2000, Wellman, Cross and Watson 2001). In the standard version of this task (the "object transfer" problem), the child is told a story in which a character's belief about the location of a target object becomes false when the object is moved without the character's knowledge. In Wimmer and Perner's original version, for instance, Maxi's mother moves the chocolate from the green to the blue cupboard while Maxi is outside playing. The children are then variously asked to report the content of the character's belief ("Where does Maxi *think* the chocolate is?"), to predict the character's action ("Where will Maxi *look* for the chocolate?"), or sometimes to explain the completed action ("Why did Maxi look for the chocolate in the *green* cupboard?"). The critical feature of a False Belief task is that the correct answers to all three of these questions – even the ones that do not specifically query a belief-content – require the child to pay attention to Maxi's belief, and not the actual location of the chocolate (Dennett 1978, Premack and Woodruff 1978). Dozens of versions of the False Belief problem have now been used, and while the precise age of success varies between children and between task versions (Wellman, Cross and Watson 2001), in general children younger than three or four years old do not correctly solve False Belief problems, but older children do.

The contentious issue is not when success on False Belief problems emerges but what such success reflects. One possibility is that children 3 or 4 years old undergo real

change in the concepts they use to reason about other minds, acquiring a previously absent representational concept of belief (Perner 1993, Flavell 1999, Wellman, Cross and Watson 2001), and producing a consequent improvement on False Belief tasks.

A second possibility is that the concept of belief is already intact in young children but is masked in the False Belief paradigm by immaturity in other capacities that are necessary for good performance on the task. Two such candidate capacities are inhibitory control and some aspects of syntactic knowledge (especially complement syntax) , both of which are correlated with False Belief task performance (Astington and Jenkins 1995, 1999, Carlson and Moses 2001, Watson, Painter and Bornstein 2001, de Villiers and Pyers 2002). Below I will review the evidence for an association between the development of reasoning about beliefs and inhibitory control or syntax. I will focus in particular on evidence for (and against) False Belief task performance as a measure of a newly acquired representational concept of belief, as opposed to simple unmasking by maturation of these other capacities.

3.1.1 Inhibitory control and belief attribution

To answer a False Belief question correctly, a child must be able to juggle two competing representations of reality (the actual state of affairs and the reality represented in the protagonist's head) and to inhibit an incorrect but compelling answer (the true location of the object). In variants of the False Belief task, the demand for inhibitory control predicts children's performance. For instance, the current location of the target object may be made less salient and thus easier to inhibit: instead of being moved to the green cupboard, Maxi's chocolate is eaten, or the actual location of the chocolate is unknown to the child (described in Wellman, Cross and Watson 2001, see also Zaitchik 1991). Children of all ages perform better on these versions of the task (Wellman et al 2001). Conversely, when the inhibitory demands are increased by changing the protagonist's motivation to a negative desire (i.e. the protagonist's desire is not to find, but to *avoid* the target object) four and even six year olds consistently fail to answer correctly (Leslie and Polizzi 1998, Leslie 2000). Finally, the inhibitory demands of False Belief tasks are not restricted to reasoning about beliefs. Four year olds have more difficulty with logically equivalent problems about non-mental false representations (e.g. false photographs or maps, Zaitchik 1990, Leslie and Thaiss 1992), and with juggling two different verbal labels of a single object (Apperly and Robinson 2002).

Nevertheless the interpretation that False Belief performance is limited by immature inhibitory control remains controversial (e.g. Perner, Stummer and Lang 1999). Wellman, Cross and Watson (2001) note that the improvement in three year olds' performance on False Belief tasks with salience manipulations reflects only a change

from below chance to chance performance, and therefore does not implicate an operational concept of belief (but see Moses 2001). They argue that the correlation between inhibitory control and False Belief performance need not reflect masking of a pre-existing competence. Instead, inhibitory control could facilitate knowledge acquisition and conceptual change in the domain of other minds, since a child who can disengage from the prepotent representation of reality may be more able to focus on and learn about mental representations (Wellman, Cross and Watson 2001, Moses 2001).

In all, the role of executive function or inhibitory control in reasoning about beliefs remains open to investigation. Is inhibitory control recruited during successful False Belief task performance? Is it recruited even more generally during all reasoning about beliefs? Is the locus of such inhibition predominantly peripheral (resolving response conflict) or cognitive (inhibiting prepotent representations)? Alternatively, if executive function contributes only to children's early learning about the mind, then we would not expect the same brain regions to be recruited when adults engage in belief attribution and inhibitory control. Below, I will examine whether neuroimaging of healthy human adults can help resolve some of these questions.

3.1.2 Language and belief attribution

A striking demonstration of the role of language in False Belief task performance comes from studies of deaf children. Deaf children of hearing parents (i.e. non-native signers) are impaired on sign language versions False Belief tasks (Peterson and Siegal 1995, de Villiers and de Villiers 1999). This deficit persists even on non-verbal (i.e. pictorial) versions of the False Belief task (de Villier and de Villiers 1999, Woolfe, Want & Siegal 2002). Deaf children of native signers show no impairment. Conversely, the quantity and quality of family talk about mental causes to which a normally developing toddler is exposed is correlated with performance on False Belief tasks two years later (e.g. Cutting and Dunn 1999).

The causal relationship between developing competences in language and "theory of mind" has been controversial, though. Some models propose that linguistic competence is a necessary precursor of theory of mind (e.g. de Villiers 2000) while others suggest that theory of mind is a necessary precursor of language development (e.g. Baron-Cohen, Leslie and Frith 1985, Happe 1992, Bloom 2000). One possibility is that early developing components of theory of mind (discussed in detail below) are necessary for some aspects of language acquisition (e.g., establishing the referents of newly heard words), whereas other aspects of language acquisition such as the syntax of complementation and the semantics of opacity are in turn necessary for the late

developing concept of belief (Malle IP). I will address the relationship between belief attribution and language here.

The specific attribute of language commonly implicated in representing another person's beliefs is the syntax and semantics of sentential complements. De Villiers (e.g. 2000, de Villiers and de Villiers 2000) has proposed that "language is the only representational system that could" support the concept of (false) beliefs, because language is "propositional, and can therefore capture falsity and embeddedness of propositions." Mental state verbs share with verbs of communication a particular syntactic structure of referentially-opaque embedded complements (the truth value of the sentence is independent of the truth value of the complement, as in "John thinks that *it is raining.*") Children's production and comprehension of this syntactic structure precedes and strongly predicts performance on both standard and non-verbal versions of the False Belief task (de Villiers and de Villiers 1999, de Villiers 2000, but see Bartsch and Wellman 1995, Ruffman et al 2003).

However, linguistic skills may correlate with reasoning about beliefs simply because language enables a child to learn about (or to learn to talk about) the mind. Conversational experience (including comprehension of embedded sentence complements) contributes to children's developing knowledge about the mind, because first-person verbal report is our dominant source of information about subjective states occurring inside someone else's head (Harris 1989, Nelson 1996). Thus, sophisticated syntax may support the development of a concept of belief, but not be recruited during reasoning with that concept.

What then is the role of language in adult reasoning about beliefs? If constructing the syntax of embedded propositional structures is necessary for all reasoning about (false) beliefs, then the same neural structures should be recruited by tasks that tap these two processes. I will address this prediction below.

3.1.3 Beyond False Belief

In all, the extensive literature on False Belief task performance in normally developing children has produced a very consistent pattern of results and many competing interpretations. Wellman et al (2001) thus concluded their meta-analysis of this literature with a plea that researchers move on to seek new and converging evidence for competing theories of the developing concept of belief. In fact, Bloom and German (2000) specify two reasons that researchers should abandon the False Belief task as the benchmark of mature belief attribution. First, there is more to passing the False Belief task than a concept of belief, as illustrated by the preceding review. Second, they point out, there is more to a concept of belief than passing the False Belief task. In fact, theory

of mind reasoning would not work, if we did not attribute to others mostly true beliefs (and mostly rational actions, Dennet 1996). For all of these reasons, the False Belief task is at best a limited tool for measuring the developing concept of belief.

So, is there evidence for the distinct emergence of a concept of belief, beyond the False Belief task? The best such evidence is Bartsch and Wellman's (1995) investigation of children's spontaneous talk about beliefs. Bartsch and Wellman distinguish mere conversational turns of phrase ("know what?") from genuine psychological references, often identified by the children's use of contrastives: sentences using "think" or "know" that contrast expectations and outcomes, fiction and reality, or differences between individuals. The first genuine references to thoughts and beliefs in all the children recorded appear around the third birthday, significantly later than genuine psychological reference to desires and emotions, but about half a year before children spontaneously explain actions in terms of beliefs (Bartsch and Wellman 1995), or pass False Belief tasks (e.g. Wellman, Cross and Watson 2001).

Evidence from both experimental tasks and spontaneous speech thus converge on a change in children's reasoning about beliefs that occurs in the third or fourth year. Is this development reflected in a specialised neural substrate for reasoning about beliefs? If so, what is the relationship between this neural substrate, and brain regions subserving inhibitory control? Language? These are questions that could be addressed using functional neuroimaging.

3.2 Neuroimaging: Belief Attribution

The first question for neuroimaging is: can we find regions of the adult human brain that show activity specifically when subjects are required to attribute beliefs to another person? Of neuroimaging studies that attempt to address this question directly, four have followed developmental psychology in using False Belief problems (verbal and non-verbal) as the definitive belief attribution task (Fletcher et al, 1995, Gallagher et al, 2000, Vogeley et al 2001, Saxe and Kanwisher, IP). One study gave subjects simple descriptions of events involving people, and instructed subjects to "try to understand their motivations, feelings and actions" (Ferstl and von Cramon 2002). Goel et al (1995) used an original task: subjects were asked to judge whether Christopher Columbus could have identified the function of a pictured object. Across these studies, blood flow increases in a consistent pattern of brain regions when subjects reason about False Beliefs or Columbus' knowledge/ignorance (see Figure 3.2): medial prefrontal cortex (BA9), temporal poles bilaterally (BA38), anterior superior temporal sulcus (BA22) and bilateral temporo-parietal junction extending into posterior superior temporal sulcus (BA39/40/22).

Could any or all of these brain regions be a specialised neural substrate for reasoning about beliefs? The reasons to be cautious with results from the False Belief task in developmental psychology apply equally to neuroimaging results. First, there is more to solving the False Belief task than a concept of belief, and second, there is more to a concept of belief than passing the False Belief task (Bloom and German 2000). In addition, "activity" in the standard subtraction methodology of neuroimaging is only as meaningful or specific as the subtracted control condition. Therefore, I propose two basic criteria for a brain region involved in the attribution of beliefs: generality and specificity. First, the candidate region must show increased activity to any stimuli that invite the attribution of beliefs, both true and false. Second, the response must be specific to belief attribution. That is, the candidate region must not show a high response to the presence of a person per se, or during reasoning about non-mental (false) representations. An important further question is whether brain regions involved in belief attribution may be distinct from those that represent other mental state concepts, such as emotion and goal, which emerge earlier in development. I will return to this third question later in the review.

Few neuroimaging studies have directly tested the neural activity associated with attributing true beliefs. However, a number of studies have included a control condition in which the protagonist's action is not based on false belief or ignorance, but on true beliefs and perceptions of the situation (Fletcher et al 1995, Gallagher et al 2000, Saxe and Kanwisher, IP). This condition may therefore have invited belief attribution, even if reasoning about mental states was not required for successful performance. Consistent with this idea, the same brain regions that showed increased response during False Belief stories generalised their activity to the stories that involved true beliefs (Fletcher et al 1995, Gallagher et al 2000)[4]. Particularly strong confirmation comes from individual-subject and ROI analyses, showing that the very same voxels in individual subjects are activated for true and false belief attribution (as described in section 2 above, Saxe and

[4] The only brain region that did not show this pattern was medial frontal cortex, which showed a significantly increased response only during False Belief stories and not during stories about actions based on true beliefs (compared with jumbled sentence controls, Fletcher et al 1995, Gallagher et al 2000). A number of authors (Gallagher et al 2000, Gallagher and Frith 2003) have concluded that medial frontal cortex is the only cortical region uniquely involved in "theory of mind". But, as we and others (e.g. Bloom and German 2000, Scholl and Leslie 2001) have argued, selective involvement in the False Belief task per se is not equivalent to selective involvement in "theory of mind". The medial prefrontal cortex may nevertheless be involved in belief attribution. Saxe and Kanwisher (IP) found that medial prefrontal cortex activity did generalise to vignettes that did not require False Belief attribution, and activity in this region has also been reported when subjects are simply instructed to try to understand a character's motivations (Ferstl and von Cramon 2002).

Kanwisher IP). Thus, the temporo-parietal junction, superior temporal sulcus and medial prefrontal cortex may all generalise to show a strong activation for both true and false belief attribution.

A candidate region specialised for belief attribution must also be shown to be specific: that is, it must not show a high response during logically similar control conditions that do not require belief attributions. One logical element of reasoning about belief and action is the need to infer invisible causal mechanisms. A second logical ingredient, more specific to False Belief stories, that must be included in a control is the notion of a false representation. Saxe and Kanwisher (submitted) therefore had subjects read stories from two control conditions. "Mechanical Inference" stories required subjects to infer the operation of an invisible mechanical force (e.g. rusting, evaporation). "False Photograph" control stories were modelled on the false photograph paradigm used by developmental psychologists (Zaitchik 1990, Leslie and Thaiss 1992). The "False Photograph" stories can be particularly closely matched to the original False Belief stories. In a typical "False Photograph" scenario, a photograph is taken of the scene (e.g. the chocolate in the Green cupboard). After the target object has been moved (e.g. to the Blue cupboard), the subject must reason about the contrasting states of affairs in the world and in the photograph, analogous to reasoning about a false belief. None of the brain regions that produced a strong response to False Belief stories showed a high response to either of these logical control conditions, consistent with the hypothesis that these regions respond specifically during attribution of beliefs, not to any false representation or hidden cause.

Finally, any brain region specialised for representing beliefs must not respond to just the simple presence of a person present in the stimulus. Most studies have included a person in a control condition (Fletcher et al 1995, Gallagher et al 2000). Even when subjects were forced to attend to the details of a physical description of a person, the responses of the temporo-parietal junction regions bilaterally and the right anterior superior temporal sulcus were no higher than a control condition using physical description of non-human objects (Saxe and Kanwisher, IP). Recruitment of these regions requires thinking about a person's beliefs, not just her appearance.

Thus a number of brain regions, including bilateral regions of the temporo-parietal junction, posterior and anterior superior temporal cortex, and temporal pole appear to fulfil at least the basic criteria for the neural substrates of attributing beliefs: generality to both true and false belief attribution, and specificity to belief attribution rather than either (a) any reasoning about people or (b) reasoning about non-mental false representations or hidden causes in general.

3.2.1 Neuroimaging: Belief attribution and inhibitory control

Given this characteristic pattern of brain activation associated with belief attribution, I can begin to address the theoretical questions raised in the first half of this section. First I consider whether brain regions associated with inhibitory control are recruited during (false) belief attribution.

Executive (or 'inhibitory') control has multiple distinct components, each of which could contribute to false belief attribution, including monitoring and detecting the conflict between competing representations or responses, selecting the correct response, and inhibiting the incorrect (possibly prepotent) response. A recent series of elegant neuroimaging studies have attempted to distinguish the neural correlates of these components (e.g. Botvinick et al 1999, Konishi et al 1999, Braver et al 2001, Milham et al 2001, Garavan et al 2002, Sylvester 2003).

Across a range of tasks, including the Eriksen flanker task[5] (Eriksen and Eriksen 1974, Botvinick et al 1999), the Stroop task[6] (Stroop 1938, Milham et al 2001), and the Go/No-Go task (Braver et al 2001), response conflict is strongly correlated with activity in anterior cingulate cortex (ACC). Braver et al (2001) found that a region of ACC is activated for any low-frequency response condition, whether it requires a response ("target detection") or the suppression of a response ("No-Go" trials). This ACC activation is significantly posterior to the paracingulate region of activity associated with False Belief problems (e.g. Gallagher et al 2000). Brain regions consistently associated with response inhibition also include dorsolateral prefrontal cortex (BA 46/9) and superior parietal lobe (BA7, e.g. Braver et al 2001, Sylvester et al 2003). Selecting the appropriate response is associated with additional activity in bilateral frontal eye fields

---

[5] In the Eriksen flanker task, subjects make a response determined by the target object in the centre of the array. On either side of the target, distractor objects are presented, which may be associated with either no response (neutral), the same response as the target (congruent), or a different response from the target (incongruent). Botvinick et al (1999) found that the response of the ACC was highest on incongruent trials which followed congruent trials, when subjects' selective attention was relatively relaxed thus provoking strong response conflict.

[6] In the Stroop task (Stroop, 1938) subjects are required to name the ink colour of a word. Response conflict increases when the word is the name of a different colour (e.g. the subject see the word "BLUE" printed in red ink, and must respond "red".) Milham et al (2001) reported that ACC activity was maximised when the word named a colour in the response set (i.e. a possible ink colour, with an assigned response), compared with trials in which the word named a colour not in the response set, consistent with the idea that ACC activity is an index of response conflict.

and intraparietal sulcus (Jiang & Kanwisher IP). None of these brain regions are among those associated with reasoning about beliefs.

Thus belief attribution – even of false beliefs – appears to rely on distinct neural systems from those responsible for response conflict, selection and inhibitory control. Consistent with this conclusion, the stories which describe a non-mental false representation – the False Photograph stories – require similar inhibitory control to False Belief stories, but do not elicit responses in regions associated with belief attribution (Saxe and Kanwisher IP). At least for adults, then, false belief attribution may not depend on inhibitory control during task performance. This is consistent with Wellman, Cross and Watson's (2001) interpretation of the developmental correlation between inhibitory control and theory of mind: namely, that inhibitory control may help children learn about the mind.

An interesting exception is the target detection paradigm. When subjects are required to respond to a low-frequency, unpredictable or unexpected target, brain activity increases in the temporo-parietal junction (Shulman et al 2002, Corbetta et al 2002, Downar et al 2002, Braver et al 2001), in a region near that activated by False Belief tasks. Again, however, no study has directly compared these two paradigms within a single experiment. Are there distinct but neighbouring sub-regions of the temporo-parietal junction involved in belief attribution and target detection? Or is there a functional relationship between these two tasks? This issue remains open for future work.

3.2.2 Neuroimaging: Belief attribution and language

As described above, De Villiers (2000) has advanced a specific, and testable, hypothesis about the relationship between language and belief attribution: that is, that language is "the only representational system that could" support the concept of a false belief, because it allows the syntactic construction of embedded sentence complements. If so, then we would predict that both belief attribution, and sentence-level syntax, recruit the same neural structures. Do they?

Studies that vary the syntactic complexity of sentences often find brain activation in and around Broca's area in left inferior frontal cortex (Caplan 2001). The location of this region of activation is inconsistent across studies, but is not in close proximity to any of the regions implicated above in belief attribution. However, two other regions are also associated with syntax in some studies: left anterior superior temporal sulcus, near the temporal pole (Friederici 2001, Vandenberghe et al 2002, Ferstl and von Cramon 2002) and posterior regions of the superior temporal sulcus, near Wernicke's area (Just 1996, Caplan 2001, Ferstl and von Cramon 2002, Roder et al 2002, Ben Shachar et al IP).

How do the regions of superior temporal sulcus that are involved in syntax compare to the regions of superior temporal sulcus implicated in belief attribution? One study has included versions of the two tasks within a single experiment. Ferstl and von Cramon (2002) scanned subjects while they read sentence pairs in one of two conditions. First, in the logic condition, subjects read two sentences describing a mechanical causal sequence, and were asked to judge whether the sequence was coherent[7]. In the second half of the experiment, subjects read sentences describing an event involving people, and were asked to try to "understand their feelings, motivations and actions". The controls for both conditions were sentences in pseudo-language. Compared to this control, the "logic" and the "theory of mind" conditions elicited increased brain activity in strikingly similar regions of posterior and anterior superior temporal sulcus and temporo-parietal junction. Does this mean that syntax and belief attribution recruit similar regions of pSTS? Unfortunately, it is not clear that Ferstl and von Cramon have isolated the parts of superior temporal sulcus responsible for syntax[8]. In their pseudo-language control condition, half of the sentences contained familiar syntax, with pseudo-words in place of content nouns. Sentences like this, with the syntax but not the content of real sentences, activate the syntax-related region of superior temporal gyrus almost as strongly as normal sentences (Roder et al 2002). Moreover, the focus of activity associated with syntax in group analyses is typically located 2 or 3 cm anterior to the focus of activity associated with belief attribution (e.g. Ben Shachar et al IP).

The relationship between the neural correlates of syntax and of belief attribution remains open to investigation, but the strong conclusion that syntax and belief attribution recruit the same brain regions seems unlikely. Neuropsychological evidence provides an

---

[7] In their paper, Ferstl and von Cramon (2002) focussed the effect of coherence. In the "logic" condition, the response of region of medial prefrontal cortex associated with false belief attribution increased significantly only when the pair of sentences was judged to be coherently connected. This may be important, because a number of the early studies of false belief tasks (e.g. Fletcher et al 1995, Gallagher et al 2000) used random unlinked sentences as their baseline control condition. Ferstl and von Cramon conclude that the medial frontal cortex is responsible generally for "the maintenance of nonautomatic cognitive processes."

[8] It is a serious challenge to explain what did account for the activation of belief attribution related regions of superior temporal sulcus in the "logic" condition of Ferstl and von Cramon (2002). Frith and Frith (2003) speculate that some of the stimuli may have invited attribution of beliefs and human actions in spite of the absence of people in the explicit descriptions. However, this seems unlikely: the "logic" stimuli were similar to the "mechanical inference" control stories used by Saxe and Kanwisher (IP) which did not elicit any response in the temporo-parietal junction or superior temporal sulcus regions. A different possibility is that the "coherence" instructions used by Ferstl and von Cramon in the 'logic' condition encouraged subjects to consider the intentions of the author of the passage.

additional hint that these two functions may be independent, at least in adulthood. Siegal and colleagues (Varley & Siegal 2000, Varley, Siegal and Want 2001) have shown that two dense aphasics, who have dramatically impaired syntactic processing following strokes, nevertheless can attribute beliefs and even pass a non-verbal False Belief task.

3.2.3 Belief attribution: Conclusions

Developmental psychology and neuroimaging studies provide converging evidence for an anatomically and functionally distinct system for belief attribution in normally developing human children and adults. Children begin to use a novel representational concept of belief around age 3, and by 4 this concept is robust enough to support successful performance on False Belief tasks. The attribution of beliefs also seems to be associated with distinct brain regions, including the medial prefrontal cortex and temporo-parietal junction. One theoretical question about the emergence of belief attribution is the extent to which it depends only on maturation of other capacities. False Belief task performance is correlated with both inhibitory control and syntax development. However, I have shown that these capacities recruit different regions of the brain, suggesting that distinct mechanisms are involved in belief attribution, inhibitory control and syntax. The developmental correlation may therefore reflect the facilitation of knowledge development about the mind for children with mature inhibitory control or linguistic skills.

The most striking dissociation in the development of theory of mind, though, is that between the late development of the concept of belief, described above, and the much earlier development of attribution of other mental states, including desires, perceptions, and emotions (e.g. Bartsch and Wellman 1995). Henry Wellman has characterised this difference as a transition from 'Desire Psychology' to 'Belief-Desire Psychology' (e.g. Wellman 2001, Bartsch and Wellman 1995). Many models of theory of mind include a similar distinction, although the precise characterisations and developmental time-courses differ (e.g. Baron-Cohen 1997, Leslie 2000, Tager-Flusberg and Sullivan 2000). In the next section, I will investigate whether a similar distinction is instantiated in the brain regions recruited by theory of mind reasoning.

4.1 Developmental Psychology: Desires, Perceptions, Emotions

Passing the False Belief task is sufficient but not necessary evidence of having a theory of mind. According to the criteria originally set out by Premack and Premack (1978), young preschoolers and even infants possess a full "theory of mind" because they can impute unobservable mental states to themselves and others, and use these mental states as a coherent framework to make predictions about behaviour. The central notion

in this earlier theory of mind seems to be a concept of desire or goal, but it also includes concepts of perception (or attention) and emotion. In this section, I will discuss each of these concepts, and then consider whether they are processed and represented in one common system or multiple distinct systems. In each case, I will discriminate between two alternative accounts of the children's competence. According to the mentalistic (rich) interpretation, young children attribute mental to states to others in order to predict and explain behaviour. According to the lean interpretation, children have simply developed (or possess innately) sensitivities to certain physical cues, like changing eye gaze or a smile, without making any mental state attributions (Povinelli 2001). The evidence in this section that children do indeed make coherent, causally interrelated mental state attributions of desires/goals, perceptions and emotions will help to establish the criteria for the neuroimaging studies that I will examine in section 4.2.

4.1.1 Desires/Goals

Children already make genuine psychological references to desires in their spontaneous speech by their second birthday; references to beliefs appear six to twelve months later (Bartsch and Wellman 1995). In the lab, two year old children respond appropriately to another person's desires, even when they differ from the child's own preferences (e.g. giving an adult experimenter more of the snack that she preferred [broccoli] rather than of the one the child liked better [crackers], Repacholi and Gopnik 1997, see also Rieffe, Terwogt et al 2001). Fifteen month olds can distinguish between the intended goal of an action, and its accidental consequences, selectively imitating the goal (Meltzoff 1995, Carpenter et al 1998). Furthermore, young children's attribution of goals is not restricted to human actors, but includes non-human agents capable of contingent interactions (Gergely & Csibra 1997, Johnson et al 2001, but see Legerstee and Barillas 2003).

Strikingly, the concept of a goal seems to be available to pre-verbal infants. Five to eight month old infants who have habituated to a reach-and-grasp motion by a human hand look longer when the object of the motion changes than when the physical path of the motion changes (Woodward 1998). That is, these infants appear to impute an unobservable mental state to the agent (a desire or goal to have the target object), and use this attribution to make a prediction about future behaviour (that the agent will continue to reach for the same object and not, for instance, towards the same location in space).

What is the nature of this early concept of desire/goal? Although the concepts of desire, goal and intention play distinct roles in adult speech and reasoning (e.g. Malle and Knobe 2001), Astington (2001) has argued plausibly that toddlers have instead an undifferentiated notion of a volitional state (or 'conation') tied to an action or an object.

Critically, the toddler does not attribute knowledge or beliefs about the apple to Anne ("Anne thinks there is an apple"), but simply uses her own knowledge of the world ("there is an apple") plus a volitional connection from Anne to the apple (approximately "Anne wants the apple", Figure 1b) to predict Anne's action (Bartsch and Wellman 1995, Wellman and Cross 2001)[9]. An undifferentiated concept of volition is in principle sufficient to support everything from the 5-month-old's simple goal-directed interpretation of reaching (Woodward 1998) to the 24–month-old's understanding that different people can have different preferences (Repacholi and Gopnik 1997). However it remains an open question whether the concept of desire/goal is essentially unchanged in the first two years of life and merely becomes more robust, or whether there is real conceptual change from a representation of a goal of hand actions in particular to a more general notion of desire.

Thus young children can attribute desires and goals long before they attribute beliefs. However, an important criterion of a theory of mind is that different mental state attributions interact in a coherent causal framework in order to allow explanations and predictions of action. In the next two subsections, I will present evidence for such interactions between the early concepts of desire/goal, perception and emotion.

4.1.2 Perceptions

Like the concept of desires/goals, a preliminary concept of perception (i.e. another person's ability to see or look at something) is available to children long before belief attribution. Verbs for seeing are in children's productive vocabulary by 26 months (Bretherton and Beeghly 1982), and children spontaneously produce embedded-clause syntax with "see" six months earlier than with "think" (Bartsch and Wellman 1995). The earliest sensitivity to others' eyes is evident before infants are 3 months old, expressed by a preference for faces with open eyes (Bakti, Baron-Cohen et al 2000) and by orientation towards the direction of gaze of a previously viewed face (Hood, Willen and Driver 1998). However, early (and later) behavioural sensitivity to another's gaze is open to both rich and lean interpretations. According to the rich interpretation, infants (or young children) understand seeing as a mental state, providing the actor with selective visual access to (and possibly knowledge about) the world. According to a lean interpretation,

---

[9] This characterisation helps to resolve the puzzle of how the concept of desire could precede, and be independent of the concept of belief. In some frameworks (e.g. Searle 1983) the concepts of desire and belief are logically identical, each composed of a proposition ("it will rain tomorrow") and an attitude (either "I want" or "I think"). However, according to this interpretation, young children do not have such a representational concept of desire. Rather, desire is conceived as a connection between a person and a real object, finessing the need for toddlers to represent a representation.

on the other hand, gaze following does not demonstrate any mentalistic understanding (see Povinelli 2001); rather, infant's tendency to follow gaze could depend on a learned association between the direction of adult gaze and interesting events in the world or on a hard-wired reflex (e.g. Corkum and Moore 1995).

When do children possess a genuinely psychological concept of perception? By early in their second year, children's concept of perception shows two hallmarks of being incorporated in their theory of mind: perception is understood to be referential, and to interact coherently with other attributed mental states, including goals and emotions. (I will return to this second point in the discussion of emotion understanding).

A referential concept of perception specifies that gaze is directed *at* something; i.e. that looking is a relationship between a person and an object. A series of recent studies show that by 12 to 14 months, infants interpret others' gaze as referential (Caron et al 2002, Brooks and Meltzoff 2002, but see Doherty and Anderson 1999). Brooks, Caron and Butler (described in Caron et al 2002) adapted Woodward's paradigm from goal directed actions to measure referential understanding of perception. Fourteen month old infants were habituated to an adult with either open or closed eyes, who turned towards one of two objects. When the position of the objects was switched at test, the infants looked longer if the adult now looked to the old location (new object) than if the adult looked to the new location (old object), but only if the adult's eyes were open during habituation, suggesting that the infants understood that looking, but not head-turning, is object-directed. Gaze following is also not restricted to human gaze. Johnson et al (1998) found that 12 month old infants would follow the 'gaze' of a faceless unfamiliar object following only 60 seconds of contingent interaction between the baby and the object, suggesting that gaze following at this age is not merely a conditioned response.

Infants use their mentalistic understanding of gaze in word learning. Dare Baldwin (1993) gave 14 and 18 month old infants one object to play with, while another object was put into a bucket in front of the experimenter. When the infant was looking at the object she was playing with, the experimenter looked into the bucket, and introduced a novel noun: "It's a blicket!" This introduces a perfect perceptual association between the novel word and the object in the infant's hand. Instead of learning this association, the infants looked up to see where the experimenter was looking. The one year olds then associated the new word "blicket" with the object that the experimenter was looking at, and not the one in their hand.

While the early concept of perception is referential, it is probably not representational. As with goals, young children may conceive of another person's

perception as a direct connection between the person and the real object in the world (without positing any internal mental representation, Figure 1). This distinction helps to explain the long delay from the (referential) understanding that perception is object-directed, around 14 months, to the (representational) notion that perception can be inaccurate or only partial and so can lead to misperceptions, which is still developing in children 4 and 5 years old (e.g. Gopnik and Astington 1988, Lalonde and Chandler 2002).

4.1.3 Emotions

Newborn infants already show sensitivity to, and synchrony with, others' emotions: they cry when other infants cry (e.g. Simner 1971) and show some discrimination of happy versus sad expressions (Field & Walden 1982). By six months, infants undeniably discriminate facial expressions (Nelson 1987, Caron, Caron and Maclean 1988), and by their first birthday, infants use parental emotional expressions, for example of fear, to guide their own actions in novel situations ('social referencing', Feinman 1992). But the discrimination of emotional expressions is open to a lean, non-mentalistic interpretation. For instance, the physical configuration of emotional expressions may constitute (either learned or innate) intrinsic rewards or punishments to the perceiver (e.g. Blair 2003). When, then, do other's emotions become incorporated in infants' emerging theory of mind? As mentioned above, one hallmark of such incorporation is the ability to form coherent combinations of attributed mental states, which emerges around 14 months.

By 14 months, infants can combine information about a person's gaze and emotion both to infer the person's goal, and to direct the infant's own actions. Phillips, Wellman, and Spelke (2002) habituated infants to an event in which an adult looked and smiled at one of two available objects, and then was shown holding the same object. At test, the adult looked and smiled either at the same (old) object or at the other (new) object, and then was shown holding the new object. Infants looked longer when the adult took the new object after gazing at the old object, suggesting that infants used gaze and emotion cues to infer the actor's subsequent goal[10]. Infants can also use emotional expressions to infer a previous goal. In a study by Tomasello, Strosberg and Akhtar (1996), an adult used a non-word to announce an intention to find an object. The adult first picked up one object with obvious disappointment and rejected it, and then picked up a second object with glee (all non-verbally). 16 month old infants learned that the

[10] In an elegant control, the authors showed that if the contingency was reversed during habituation (i.e. the adult looked at one object, and then was shown with the other one) infants were not able to generalise this pattern to the test trials. Thus, infants were not simply learning a pattern of contingencies during the experiment.

novel word referred to the object of the positive emotion, suggesting that infants know that happiness results from goal-fulfilment. Finally Moses, Baldwin, Rosicky and Tidball (2001) adapted Baldwin's (1993) word learning paradigm to show than infants use the gaze direction of an adult to determine the referent of an emotional message. When an adult made a negative emotional noise, 12 and 18 month olds looked up from the novel object in their hand to determine the gaze direction of the adult. Infants subsequently avoided only the object that had been the focus of the adult's gaze.

The causal relation between attributions of emotions and desires is also apparent once children begin to speak. Around 24 months, children begin to spontaneously talk about the causes of their own emotions (Bretherton, Fritz, et al 1986, Wellman, Banerjee, Harris and Sinclair 1995), especially the relationship between frustrated desires and negative mental states (Dunn and Brown 2001, Lagattuta & Wellman 2001). In the lab, 2 year olds told about a boy who wanted a puppy and got one choose a happy face to show how the boy would feel, but the same children choose a sad face if the boy had wanted a bunny (Wellman and Woolley 1990).

Of course, as with concepts of perception and desire, children's understanding of emotions continues to develop after age 2. Complex social emotions, like pride, embarrassment and guilt begin to be correctly attributed between the ages of 5 and 14 (e.g. Berti, Garattoni & Venturini 2000,). But the framework for attributing basic emotions and their causal relations with desires and perception appears to be already intact in the second year of life.

4.1.4 Common or distinct mechanisms?

In all, in their second and third year, toddlers reason productively and coherently about action, using basic concepts of (and causal relations between) three kinds of mental states: desires or goals, perceptions, and emotions. This early theory of mind seems to emerge significantly earlier than the reasoning about beliefs, described in the first half of this review.

An open question concerns the extent to which attributions of desires, perceptions and emotions rely on distinct or common functional or anatomical substrates. For instance Alan Leslie (1995), in his hierarchical model of theory of mind development, lumps perceptions and goals together as "actional properties", represented by a stage he calls the Theory of Mind Module 1. Actional properties are those that let an agent "act in pursuit of goals, react to the environment and interact with each other." Consistent with this proposal, in a longitudinal study of 9 to 15 month olds, Carpenter et al (1998) found that the emergence of attentional engagement and gaze following (attribution of perception) and imitation of novel actions (goal attribution) were positively correlated

with each other, and uncorrelated with concurrent non-social developments such as object permanence. The authors conclude that a mentalistic understanding of gaze and action are "two instances of the same underlying phenomenon". Simon Baron-Cohen (1994), on the other hand, has divided this domain into two distinct components. The 'Intentionality Detector' represents behaviour in terms of goals, while the 'Eye Direction Detector' detects eyes and represents the direction as the Agent 'seeing'. Neither model explicitly includes a mechanism for emotion attribution.

4.2 Neuroimaging: Desires, Perceptions and Emotions

The above review suggests three central questions that could be addressed using neuroimaging. First, reasoning about beliefs develops later than, and may depend upon (e.g. Pellicano and Rhodes 2003), an earlier theory of mind that includes attribution of desires, perceptions and emotion. Does the later emerging competence colonise the same neural systems that underpin earlier reasoning? If so, we would predict that attributions of desires, for instance, would recruit the same brain regions identified above as involved in belief attribution. If, on the other hand, reasoning about beliefs draws on distinct systems or abilities, then desire attribution should not produce activity in regions associated with belief attribution, and may recruit a distinct set of brain regions.

Second, the models of early theory of mind proposed by Leslie (1995) and Baron-Cohen (1994) disagree about whether attributions of perception and goals are the province of one common or two distinct modules. A preliminary approach to resolving this controversy is to ask: do these functions recruit the same or different brain regions in adults?

Finally, what is the relationship between the attribution of emotions and of other mental states?

4.2.1 Neuroimaging: Attributing Desires and Goals

In functional neuroimaging studies, the attribution of desires, goals and intentions to another person has been investigated in three basic paradigms. Studies of the first kind invite mental state attribution to a fictional character using vignettes, cartoons, or animations similar to the stimuli used in studies of belief attribution (Fletcher et al 1995, Gallagher et al 2000, Castelli et al 2000, Castelli et al 2002, Brunet et al, Saxe and Kanwisher IP, Schultz et al 2003). In a second, related set of studies, subjects engage in a simple game, purportedly either with an unseen agent (presumably inviting goal attributions) or with a computer (discouraging goal attributions, McCabe et al 2001, Gallagher et al 2002). Finally, in the third kind of study, subjects watch and interpret a (video of a) simple goal-directed action by a human actor (Chaminade, Meltzoff and

Decety 2001, Decety et al 2002, Koski et al 2002, Zacks et al 2001, Saxe et al submitted). Below I consider each of these three sets of studies sequentially. Unfortunately, very few of these studies also included a task designed to elicit belief attribution, and to our knowledge only one study to date explicitly aimed to contrast brain regions involved in the attribution of different kinds of mental states (Saxe et al submitted).

Vignettes, cartoons and animations that depict or suggest a character's goals, intentions or desires are typically correlated with moderately increased activity (compared to scrambled or non-social controls) in the brain regions associated with belief attributions, including medial prefrontal cortex and posterior superior temporal sulcus (Gallagher et al 2000, Castelli et al 2000, Brunet et al, Saxe and Kanwisher IP, Buccino et al 2001, Schultz et al 2003). For instance, stories describing a character's desires elicited significantly more activity in these regions than physical descriptions of a person, but significantly less than stories describing false beliefs (Saxe and Kanwisher IP). The most important weakness of this evidence is that none of the stimuli in these studies were designed to exclude belief attribution. The intermediate activity in all of these studies may reflect weak but consistent activation of these regions during desire/goal attribution, but it may equally reflect subjects' occasional spontaneous belief attribution in response to these stimuli. Future studies are needed in which vignettes about different mental states are explicitly contrasted.

Games provide an appealing paradigm for investigations of goal attribution, because they allow the stimuli to be exactly matched during goal-attribution and no-goal conditions. Two PET studies using this logic found increased activity in regions of medial prefrontal cortex when the opponent was (purportedly) a human, compared with computer-opponent trials (McCabe et al 2001, Gallagher et al 2002). To isolate goal attribution in particular, this contrast is too broad, though. Differences in brain activity could reflect belief attribution as above, or just the felt presence of a human opponent (see description in Gallagher et al 2002; Saxe and Kanwisher [IP] reported that a region of medial prefrontal cortex responded more to any story containing a person, including physical descriptions, than to non-human control stories). No studies have directly compared belief and goal attribution within the context of a game.

Many studies have investigated the neural correlates of observing human body movements. Here I concentrate on the subset of these studies that explicitly address the perception of intentional or goal-directed action. One context in which representation of goal-directed action has been investigated is imitation (Chaminade, Meltzoff and Decety 2001, Decety et al 2002, Koski et al 2002). For example, Koski et al (2002) had subjects imitate simple index finger movements, viewed either with or without target dots (the target dots presumably made the finger movement appear goal-directed). The presence of

the goal produced increased activation in lateral inferior frontal cortex bilaterally (Broca's area, also observed by Buccino et al 2001 for object-directed versus mimed movements of hands and mouths). Furthermore, transient disruption of Broca's area by TMS (Transcranial Magnetic Stimulation) interfered with imitation, but not with cued execution, of goal-directed finger movements (Heiser et al 2003). Broca's area is not one of the regions associated with belief attribution, providing preliminary evidence that at least very simple goals may be attributed using different brain regions from those involved in reasoning about beliefs.

Finally, two studies have investigated brain regions associated with the segmentation or interpretation of whole body actions. Zacks et al (2001) looked for activity correlated with event-boundaries in, or transitions between sub-goals of, a complex goal-directed action (e.g. cleaning the kitchen). Saxe et al (submitted) varied the structure of a simple intentional action (walking across a room) by having the actor unexpectedly pause for a few seconds behind a large bookcase, perhaps requiring subjects to reformulate their interpretation of the action (the occlusion manipulation allowed the two conditions to be matched for average visual information, including biological motion). Both studies report activity related to action segmentation in right posterior superior temporal sulcus. Saxe et al (submitted) further established that this activation was specific to intentional actions, since the same pattern was not observed for passively occluded people, and generalised to other stimuli, since the same region showed a high response to two-dimensional animations portraying goal-directed 'actions' compared with rapid rigid rotation. Activity in the same vicinity was also reported by Decety et al (2002) when subjects viewed another person's hands performing an action similar to the action that they were concurrently executing, compared with viewing their own hands performing that action.

Is the region of posterior superior temporal sulcus that is involved in the analysis of intentional actionthe same as the nearby region associated with the attribution of beliefs? Saxe et al (submitted) argue that these two regions are distinct. The regions showing increased activity during False Belief stories (dubbed the TPJ-M) and during the paused walking action (dubbed the pSTS-VA, for Visual analysis of Action) did not overlap anatomically in individual subjects, and showed strikingly different functional profiles in ROI analyses.

To summarise, the evidence is inconclusive as to whether interpreting the goal or intention of an observed action draws on distinct brain regions from those involved in assigning beliefs to the actor. Evidence for distinct brain regions for attributing goals comes from studies using videos of simple actions (e.g. Koski et al 2002, Heiser et al 2003, Saxe et al submitted). Weaker evidence for common regions recruited during both

belief and desire attributions comes from studies using vignettes, cartoons, animations and games, although all of these studies have tended to confound attributions of beliefs and desires (Fletcher et al 1995, Gallagher et al 2000, Castelli et al 2000, Castelli et al 2002, Brunet et al, Saxe and Kanwisher IP, McCabe et al 2001, Gallagher et al 2002). One possible resolution is that the brain contains distinct representations of goals (restricted to simple, visible motor actions), similar to the notion of goal available to 5 – 8 month olds, and desires (applicable more generally), available after the first birthday. This speculation awaits testing in future work.

To summarise, brain regions involved in representing goal-directed action (including pSTS and Broca,s area) are distinct from the brain regions associated with belief attribution (including the TPJ and medial prefrontal cortex). Thus brain activation patterns are consistent with developmental psychology in suggesting distinct mechanisms for attributing goals and beliefs. As summarised above, an understanding of goal-directed action is available to even very young infants, while belief attribution does not emerge until three years later. The critical, and unresolved, question remains the place of desires in this scheme. Young toddlers talk about desires long before they talk about beliefs (Bartsch and Wellman 1995), and seem to conceive of desire and goals similarly, as direct volitional connections to the world (Astington 2001), suggesting that desire attribution should be similar to goal attribution and distinct from belief attribution. For adults, on the other hand, belief and desire attribution may be simply inseparable. In neuroimaging studies using vignettes, cartoons, animations and games, attributions of desires and beliefs seems to elicit activity in the same set of regions (although beliefs and desires were always confounded in the stimuli, Fletcher et al 1995, Gallagher et al 2000, Castelli et al 2000, Castelli et al 2002, Brunet et al, Saxe and Kanwisher IP, McCabe et al 2001, Gallagher et al 2002). Alternatively, desires may be reanalysed as representational as part of the conceptual transition to a representational treatment of belief. In adult folk theory, we desire a particular apple under a particular description, e.g. representing it as food that tastes good. Further work is necessary to determine whether the correct division of "natural kinds" in the adult brain places desires with goals or with beliefs.

4.2.2 Neuroimaging: Attributing Perception

In developmental psychology, a controversy remains over whether the attribution of goals and of perceptions relies on a single (e.g. Leslie 1995) or two distinct systems (e.g. Baron-Cohen 1997). One way to address this question is to ask whether attributions of goals and of perception recruit the same or distinct brain regions in adults.

Activity in the (right) posterior superior temporal sulcus (pSTS) is associated with perception of the gaze of a face: the response in this region is higher for moving eyes

than for a moving checkerboard (Puce et al 1998) or for a change in facial identity (Haxby et al 2002), for open (direct or averted) eyes than for closed eyes (Wicker et al 1998) and when subjects attend to the gaze rather than the identity of faces (Hoffman and Haxby 2000). However, as described above for the developmental studies, a sensitivity to gaze is not sufficient to indicate the attribution of perception. Perception is referential, gaze directed towards something. In an elegant recent study, Pelphrey et al (2003) showed that the response of the right pSTS to moving eyes is modulated by interaction with a target – a small checkerboard to the right or the left of the character's face. When the character quickly moved his eyes away from the target (incongruent) instead of towards the target (congruent) the response of the pSTS was sustained for many seconds longer. Pelphrey et al suggest that on incongruent trials "the observer's expectation is violated and activity in the STS region is prolonged – perhaps related to a reformulation of an expectation."

What is the relationship between regions of the pSTS[11] associated with perception of gaze changes and of other intentional actions? To our knowledge, no single study has combined target-directed directed hand or body actions and gaze changes. Puce et al (1998) reported that eye and mouth movements (with no goal) elicited activity in the same part of pSTS, but other studies using hand and body actions appeared to produce activity more lateral and anterior in the STS. On the other hand, centres of activity reported by Pelphrey et al (2003) for unexpected gaze change and by Saxe et al (submitted) for unexpected action changes are both anatomically and conceptually similar. A further suggestion of combined neural representation comes from the discovery of neurones in the anterior superior temporal sulcus of macaque monkeys that show increased response to target-directed hand actions only when the actor's gaze is directed towards the action (Jellema et al 2000). Neuroimaging work along these lines may help to determine whether attributions of referential gaze and goal directed actions rely on common or distinct brain regions, consistent with Carpenter et al's (1998) conclusion described above that these two behaviours reflect "the same underlying phenomenon."

[11] Eye gaze – especially averted gaze - was also associated with activity in medial prefrontal cortex, compared with eyes looking down or closed (Calder et al 2002). These authors point out the anatomical similarity between this activation and the region of medial prefrontal cortex associated with False Belief task performance (e.g. Gallagher et al 2000) and suggest that gaze information may recruit activity across the whole network of brain regions associated with theory of mind.

4.2.3 Neuroimaging: Attributing Emotions

Investigations of emotion have tended to proceed separately from investigations of theory of mind (but see Terwogt and Stegge 1998). Nevertheless, emotions can be attributed to others and are causally interrelated with other mental state attributions: fulfilled goals cause happiness, the object of fear can be identified by gaze direction, etc.

An overview of the neural systems associated with the perception, experience and function of emotion is available from many recent reviews (Blair 2003, Hamann 2003, Preston and de Waal 2002, Adolps 2002a, Adolphs 2002b, Canli and Amin 2002, Cardinal Parkinson et al 2002, Morris 2002, Hoffman Haxby and Gobbini 2002, LeDoux 2000). Facial emotional expressions (usually compared with neutral faces) are associated with activity in a number of different brain regions, including extrastriate cortex, right parietal cortex, right fusiform gyrus, orbitofrontal cortices, amygdala, insula and basal ganglia (Adolphs 2002b).

For the purposes of this chapter, I am restricted to the narrower question of brain regions recruited during the attribution of emotion to another person and the integration of emotions in a theory of mind. Thus we would like to distinguish neural responses to facial expressions that reflect emotion attribution, from those that reflect, for example, threat detection (e.g. Adolphs and Tranel 2000), the intrinsic reward value of a facial expression (e.g. Blair 2003), resolution of environmental ambiguity (e.g. Whalen 1999) or the initiation of behavioural withdrawal (e.g. Anderson, Spencer, Fulbright and Phelps 2000) – all of which may be independently correlated with the perception of facial expressions. We may also wish to distinguish the attribution of an emotion ("she is feeling sad") from simple emotional contagion ("this makes me feel sad"), although is it controversial whether these are indeed distinct.

One way to look for brain regions involved in the attribution of emotion may be to use descriptions of personal emotional experiences. Along these lines, Decety and Chaminade (2003) had subjects watch videos in which actors recounted experiences (in the first person) that were either sad or neutral in content. The actors' facial and emotional expressions were also manipulated to be either happy, sad or neutral. The sad narrative content, irrespective of emotional expression, led to increased neural activity in regions associated both with negative emotions (e.g. the amygdala) and in regions associated with belief attribution (e.g. the temporal pole). Interestingly, attribution of emotion was also associated with activity in left lateral inferior frontal cortex (near Broca's area), which was associated the representation of goals in the imitation paradigm (Koski et al 2002). Two distinct functional patterns were observed. In the anterior part of this region (the pars orbitalis) the neural response was high during stories with sad

negative content, regardless of the actor's expression. However, the more dorsal part (the pars opercularis) showed a high response to emotional expression (happy or sad) independent of narrative content. This pattern of results is suggestive of a dissociation between perception of facial emotion, and understanding of emotional content in speech. The relationship between representations of goals and of emotions in lateral inferior cortex requires further investigation.

A second way to identify the neural correlates of emotion attribution may be to look for interactions between attributions of emotions and of other mental states. Wicker et al (2003) investigated whether specific brain regions are sensitive to the interaction between emotional expression and gaze (specifically, direct versus averted gaze). They reported that the response of a single region in the right anterior superior temporal gyrus was highest during emotional expressions directed at the subject, compared with direct neutral gaze, and the interaction with averted gaze was significant. However these results are somewhat hard to interpret, since this region has not been previously associated either with the perception of gaze, or with the attribution of emotion. It is not possible to compare the brain regions recruited during attribution of perception and emotion in this study, since all conditions included the same gaze shift. Again, this provides no clear evidence that either common or distinct brain regions are involved in the attribution of emotion and perception.

Finally, we might ask whether brain regions identified previously in this review as associated with the attribution of beliefs, goals, or perception are also associated with perception of facial emotion. One such candidate region is the posterior STS[12]. Regions of posterior STS have been associated with representations of action (Decety et al 2002,

---

[12] Another region worth considering in this context is the right fusiform gyrus. A number of studies have reported modulation of the right fusiform gyrus by facial emotional expression (greater response to emotional that neutral faces, e.g. Veilleumier et al 2001, Halgren et al 2000). Geday et al (2003) found that posterior right fusiform gyrus activity was greater for pictures of emotional complex social scenes than for neutral counterparts. These stimuli were specifically designed to be oriented away from the observer, in order to limit the possibility of direct threat or reward to the subject. Geday et al therefore claim that their task selectively recruited regions involved in attributing emotions to others. The right fusiform gyrus has also been implicated when subjects view animations of social interactions designed to elicit mental state attributions (e.g. 'seduction', 'bullying': Castelli et al 2000, Schultz et al 2003). However, it is unclear whether fusiform gyrus activity associated with the animations reflects attributions of other mental states like desires or perceptions, or simply of emotions, since these were confounded in both studies. A further open question (cf. Schultz et al 2003) is the relationship between regions of the fusiform gyrus involved in the attribution of emotions, and the sub-region of the fusiform gyrus known as the Fusiform Face Area, or FFA (Kanwisher et al 1997) which shows a greater response for faces than all other familiar object classes.

Saxe et al submitted) and of perception (Pelphrey et al 2003). Both of these studies found increased activity in the right posterior STS when subjects' expectations about human behaviour were violated. Suggestively, Decety and Chaminade (2003) find that a nearby region produced a high response when an actor recounted a negative personal experience using positive emotional facial and vocal expression. Such incongruence between narrative content and affect may constitute a violation of expectations about behaviour in the emotional domain. (For further evidence of an association between posterior STS and emotion attribution see Narumoto et al 2001). Future work should aim to determine whether these regions of the STS reflect a common neural mechanism for the attribution of goals, perceptions and emotions, or whether neighbouring but distinct sub-regions are independently responsible for each of these functions.

In all the question of the relationship between the attribution of emotion and other components of theory of mind remains unanswered. More work is needed on both psychological and anatomical commonalities between these two critical components of understanding others.

4.2.4 Summary: Desires, Perceptions and Emotions

Evidence from developmental psychology unequivocally supports distinct psychological mechanisms for attributing desires/goals, perceptions and emotions to others (the early-developing theory of mind) from those responsible for attributing beliefs. The results of the neuroimaging studies reviewed above suggest a similar division between brain regions. Videos of simple goal directed action elicit activity in a region of posterior STS that is distinct from the nearby temporoparietal junction region (TPJ-M) associated with belief attribution (Decety et al 2002, Saxe et al submitted). A similar region of posterior STS has also been implicated in attributing perception (Pelphrey et al 2003) and even possibly emotion (Narumoto et al 2001, Decety and Chaminade 2003). Future studies are necessary to determine whether this might indeed reflect a single underlying mechanism for representing all of the (so-called 'actional') properties of people that let them act in pursuit of goals, react to the environment and interact with each other (Leslie 1995).

5. Conclusions

Substantial behavioral evidence indicates that understanding other minds follows a characteristic developmental trajectory, beginning with the early appearance (in the first 2 years of life) of a system for reasoning about other people's goals, perceptions, and emotions, and the later development (around 4 years of age) of a system for representing the contents of other people's beliefs. Here I asked whether neuroimaging research in

adults has or can contribute to theoretical debates about theory of mind that have arisen from the developmental literature. I argue that in several instances, the neuroimaging literature already provides important constraints on these debates.

First, neuroimaging has identified brain regions that are selectively engaged when people reason about the contents of other people's beliefs. This finding strengthens arguments that theory of mind constitutes a "special" domain of cognition with its own domain-specific processing machinery. Second, the brain regions associated with belief attribution appear to be distinct from other regions engaged when people reason about other people's goals, suggesting that the two stages of development result from the appearance of two distinct mechanisms, rather than the gradual enrichment of a single mechanism. Third, the brain regions associated with belief attribution appear to be distinct from those engaged in inhibitory control and from those engaged in syntactic processing. This finding argues against the hypothesis that these other functions are necessarily engaged when attributing beliefs. Instead, the neuroimaging data suggest that the reported correlations between the development of theory of mind and both inhibitory control and syntactic processing may reflect the requirement of these systems for learning about beliefs.

While these contributions from neuroimaging are substantial, they leave many other important questions unresolved. First, if we can engage reasoning about desires without engaging reasoning about beliefs, will we still see activation of brain regions associated with belief attribution, suggesting common mechanisms, or will we fail to engage the same regions, consistent with a real dissociation between reasoning about beliefs and desires? Second, are the early-developing abilities to understand other people's goals, perceptions, and actions based on a single system, or several distinct systems? Third, when strict individual-subjects analyses are applied as described in Section 2, are distinct but neighbouring sub-regions of the temporo-parietal junction involved in belief attribution and target detection? If instead these tasks engage overlapping regions, what common process might explain that overlap? Finally, I am hopeful that neuroimaging can also address other important aspects of understanding other minds that I have not had space to address here, such as the perception and neural representation of agency (Johnson 2000, Tremoulet and Feldman 2000, Csibra 2003, Johnson 2003; Ruby & Decety 2001, Farrer et al 2002, 2003), the relationship between action perception and action planning (Wolpert, Doya & Kawato 2003, Rizzolatti, Fogassi & Gallese 2001), the related problem of whether theory of mind is implemented as a pseudo-scientific theory or as a simulation (Goldman & Gallese 1998, Bartsch 2002, Stich & Nichols 1998, Nichols et 1996), the relationship between moral cognition and emotion (Greene and Haidt 2003, Moll et al 2002), and the relationship between

attribution of enduring traits (like personality) and transient states of a person (like emotions and goals, e.g. Winston 2002, Heberlein, Adolphs et al submitted).

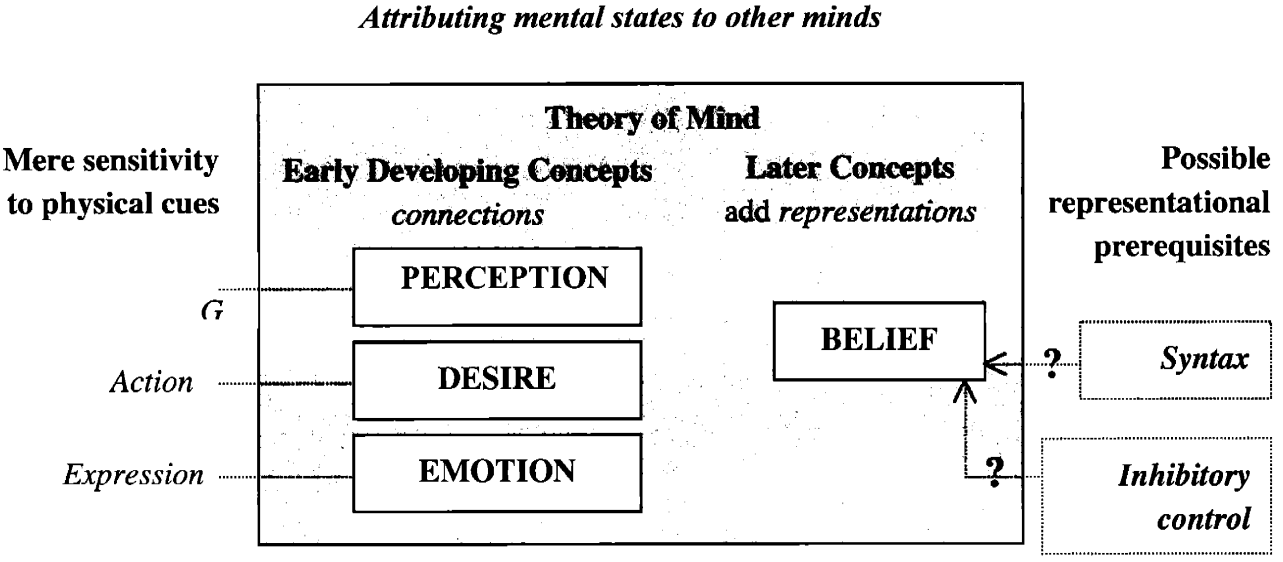
In sum, I am optimistic that neuroimaging data can help to answer fundamental questions emerging from developmental psychology about our system for reasoning about other people. These contributions are clearest for questions about the basic achitecture of the system for understanding other minds: what are its fundamental components? However, as argued in detail in Section 2, neuroimaging can make a real contribution toward answering these questions only if we uphold strict standards concerning the way the data are analyzed and the kinds of inferences we draw from them.

Figure Legends

6. Schematic representation of the two principle stages in the development of theory of mind, and the central theoretical questions discussed in this review. (a, left) Toddlers reason about other minds with a limited repertoire of mental state concepts, including desire/goal, perception and emotion. For both developmental psychology and neuorimaging, it is critical to distinguish attribution of desires, perceptions and emotions (the "rich" or "mentalistic" interpretation) from mere behavioural sensitivity to the associated physical cues (including human body motions, gaze direction and emotional expressions, the "lean" interpretation). Two important characteristics of mentalistic attribution are reference (mental states are about objects or situations) and coherence (different mental states attributions interact causally and systematically). It is an open question, addressed in this review, whether these three mental state concepts are attributed using one common mechanism or multiple distinct mechanisms. (a, right) Starting around age 3 or 4 young children include a concept of belief in their reasoning about other minds. A central debate in developmental psychology concerns whether this later development reflects true conceptual change in the child's theory of mind, or simply the maturation of other capacities which are necessary for reasoning about beliefs. Two prominent candidate capacities are syntax and inhibitory control. Evidence reviewed here suggests that representing others' beliefs does not recruit the same brain mechanisms as either syntax or inhibitory control. (b) What then is the nature of the conceptual change between a toddler's theory of mind, and the later theory which incorporates attributions of belief? One possibility is that the toddler lacks the notion of a representational mental state, and instead conceives of mental relations between people and the world as direct connections. We can think of a connection as something like gravity or resonance to affordances in the environment, leaving no possibility for error or misperception. Thus goals, perceptions and emotions may at first be understood as referential (about an object or situation) but not as representational (requiring an independent representation of the object or situation in the mind of the actor). The terms "connection" versus "representation" were used to characterise this developmental change by Flavell (1988), but a similar distinction is included in the theories of Perner (1991) and Wellman (1990). The figure is adapted with permission from Wellman and Bartsch (1995).

7. Schematic representation of brain regions associated with the attribution of mental states: beliefs, desires, perceptions and/or emotions.

Figure 1. (a)


Attributing mental states to other minds
Theory of Mind
Mere sensitivity to physical cues
Early Developing Concepts
connections
Later Concepts
add representations
Possible representational prerequisites
PERCEPTION
Gr
BELIEF
Syntax
Action
DESIRE
Expression
EMOTION
Inhibitory control

(b)

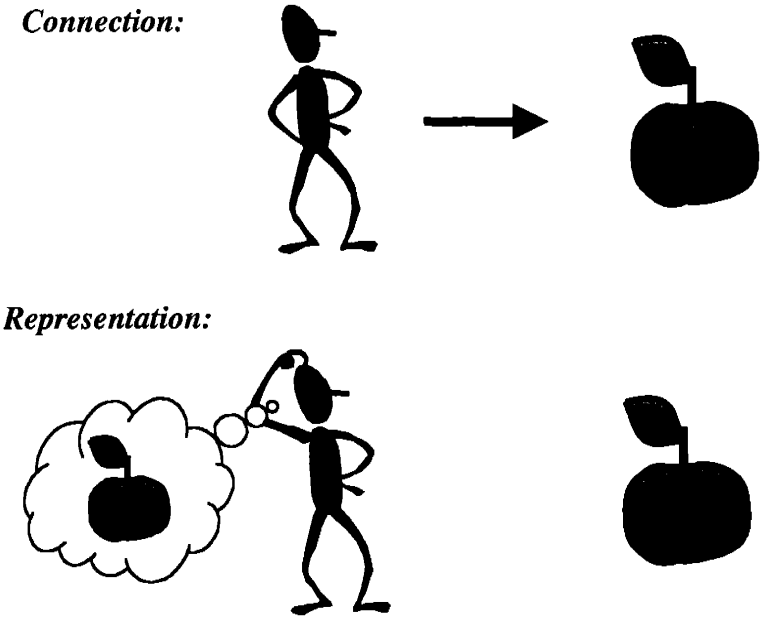
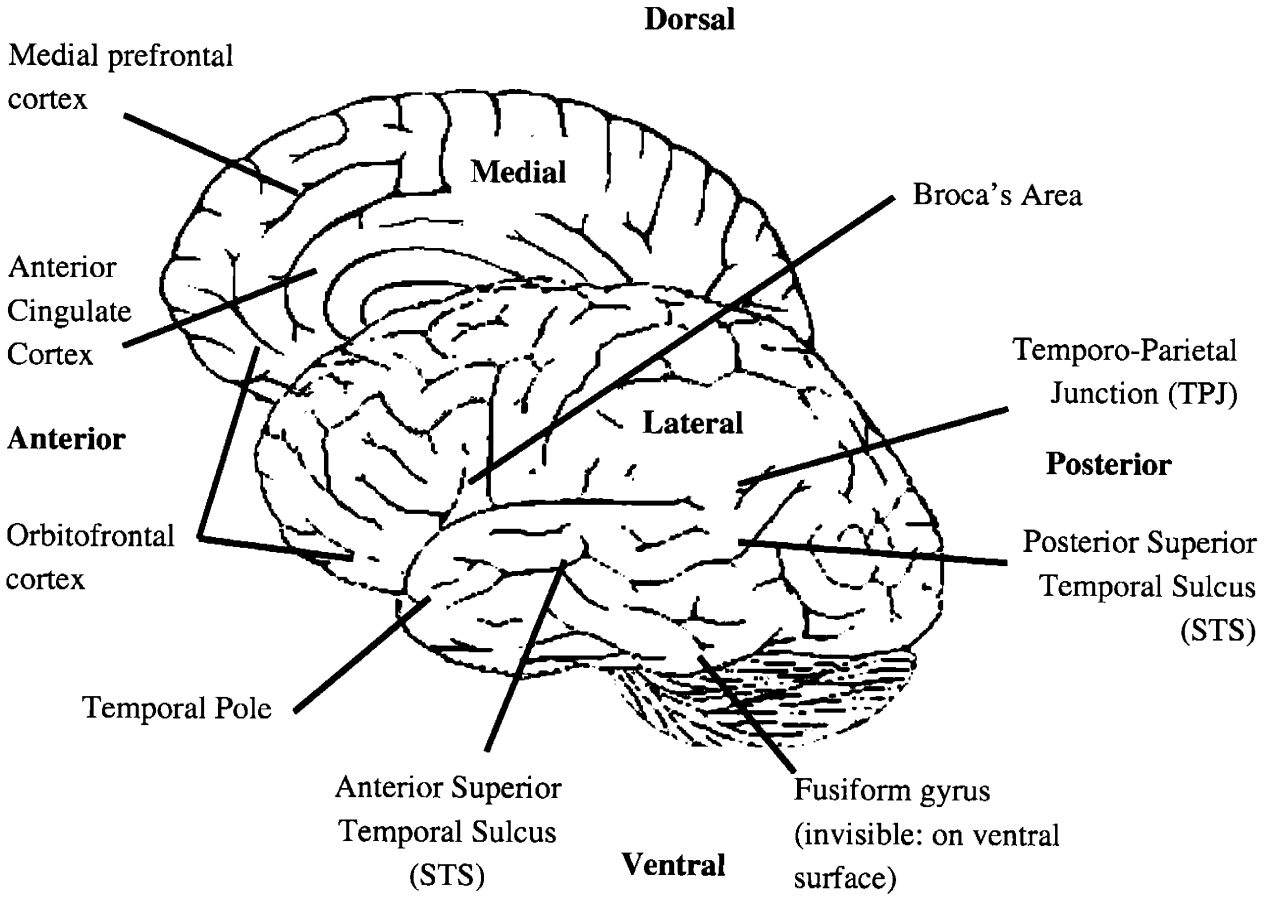

Connection:
Representation:

Figure 2.


Dorsal
Medial prefrontal cortex
Medial
Broca's Area
Anterior Cingulate Cortex
Temporo-Parietal Junction (TPJ)
Anterior
Lateral
Posterior
Orbitofrontal cortex
Posterior Superior Temporal Sulcus (STS)
Temporal Pole
Anterior Superior Temporal Sulcus (STS)
Ventral
Fusiform gyrus (invisible: on ventral surface)

References

Adolphs R. 2001. The neurobiology of social cognition. *Curr. Opin. Neurobiol.* 11: 231-9

Adolphs R. 2002. Neural systems for recognizing emotion. *Curr. Opin. Neurobiol.* 12: 169-77

Adolphs R. 2003. Cognitive neuroscience of human social behaviour. *Nat. Rev. Neurosci.* 4: 165-78

Adolphs R, Tranel D. 2000. Emotion recognition and the human amygdala. In *The Amygdala: AFunctional Analysis*, ed. JP Aggleton, pp. 587-630. New York: Oxford University Press

Allison T, Puce A, McCarthy G. 2000. Social perception from visual cues: role of the STS region.*Trends Cogn. Sci.* 4: 267-78

Anderson AK, Spencer DD, Fulbright RK, Phelps EA. 2000. Contribution of the anteromedialtemporal lobes to the evaluation of facial emotion. *Neuropsychology* 14: 526-36

Apperly IA, Robinson EJ. 2002. Five year olds' handling of reference and description in thedomains of language and mental representation. *J. Experimental Child Psychology* 83: 53-75

Astington JW. 2001. The future of theory-of-mind research: understanding motivational states, the role of language, and real-world consequences. *Child Dev.* 72: 685-7.

Astington JW. 2001. The paradox of intention: assessing children's metarepresentational understanding. In *Intentions and Intentionality: Foundations of Social Cognition*, ed. BFMalle, LJ Moses, DA Baldwin. Cambridge MA: The MIT Press

Astington JW, Jenkins JM. 1995. Theory of Mind development and social understanding. *Cognition and Emotion* 9: 151-65

Astington JW, Jenkins JM. 1999. A longitudinal study of the relation between language andtheory of mind development. *Dev. Psychol.* 35: 1311-20

Bakti A, Baron-Cohen S, Wheelwright A, Connellan J, Ahluwalia J. 2000. Is there an innate gaze module? Evidence from human neonates. *Infant Beh. Dev.* 23

Baldwin DA. 1993. Infants' ability to consult the speaker for clues to word reference. *J. Child Lang.* 20: 395-418

Baldwin DA, Baird JA. 2001. Discerning intentions in dyanmic human action. *Trends Cogn. Sci.* 5: 171-78

Baron-Cohen S. 1997. *Mindblindness: An Essay on Autism and Theory of Mind.* Cambridge MA:The MIT Press. 171 pp.

Baron-Cohen S, Leslie A, Frith U. 1985. Does the autistic child have a theory of mind? *Cognition* 21: 37-46

Baron-Cohen S, Tager-Flusberg H, Cohen DJ, eds. 2000. *Understanding Other*

Minds. Oxford: Oxford University Press

Bartsch K. 2002. The role of experience in children's developing folk epistemology: review and analysis from the theory-theory perspective. *New Ideas in Psychology* 20: 145-61

Bartsch K, Wellman HM. 1995. *Children Talk About the Mind*. Oxford: Oxford University Press

Berthoz S, Artiges E, Van De Moortele PF, Poline JB, Rouquette S, et al. 2002. Effect of impaired recognition and expression of emotions on frontocingulate cortices: an fMRI study of men with alexithymia. *Am. J. Psychiatry* 159: 961-7.

Blair RJ. 2003. Facial expressions, their communicatory functions and neuro-cognitive substrates. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 358: 561-72

Blakemore SJ, Decety J. 2001. From the perception of action to the understanding of intention. *Nat. Rev. Neurosci.* 2: 561-7

Bloom P. 2000. *How Children Learn the Meanings of Words*. Cambridge, MA: The MIT Press. 300 pp.

Bloom P, German TP. 2000. Two reasons to abandon the false belief task as a test of theory of mind. *Cognition* 77: B25-31.

Born P, Leth H, Miranda MJ, Rostrup E et al. 1998. Visual activation in infants and young children studied by functional magnetic resonance imaging. *Pediatr Res* 44: 578-83

Botvinick M, Nystrom LE, Fissell K, Carter CS, Cohen JD. 1999. Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature* 402: 179-81

Bretherton I, Beeghly M. 1982. Talking about internal states: the acquisition of an explicit theory of mind. *Dev. Psychol.* 18: 906-21

Bretherton I, Fritz J, Zahn-Waxler C, Ridgeway D. 1986. Learning to talk about emotion: a functionalist perspective. *Child Dev.* 57: 529-48

Brooks R, Meltzoff AN. 2002. The importance of eyes: how infants interpret adult looking behavior. *Dev. Psychol.* 38: 958-66

Brunet E, Sarfati Y, Hardy-Bayle MC, Decety J. 2000. A PET investigation of the attribution of intentions with a nonverbal task. *Neuroimage* 11: 157-66.

Buccino G, Binkofski F, Fink GR, Fadiga L, Fogassi L, et al. 2001. Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *Eur. J. Neurosci.* 13:400-4.

Buitelaar JK, van der Wees M. 1997. Are deficits in the decoding of affective cues and in mentalizing abilities independent? *J. Autism Dev. Disord.* 27: 539-56.

Burgund ED, Kang HC, Kelly JE, Buckner RL, Snyder AZ et al. 2002 The feasibility of a common stereotactic space for children and adults in fMRI studies of development. *Neuroimage* 17: 184-200

Calder AJ, Lawrence AD, Keane J, Scott SK, Owen AM, et al. 2002. Reading the mind from eye gaze. *Neuropsychologia* 40: 1129-38

Canli T, Amin Z. 2002. Neuroimaging of emotion and personality: scientific evidence and ethical considerations. *Brain Cogn.* 50: 414-31

Caplan D. 2001. Functional neuroimaging studies of syntactic processing. *J. Psycholinguist. Res.* 30: 297-320

Cardinal RN, Parkinson JA, Hall J, Everitt BJ. 2002. Emotion and motivation: the role of the amygdala, ventral striatum and prefrontal cortex. *Neuroscience and Biobehavioural Reviews* 26: 321-52

Carlson SM, Moses LJ. 2001. Individual differences in inhibitory control and children's theory of mind. *Child Dev.* 72: 1032-53

Caron AJ, Butler S, Brooks R. 2002. Gaze following at 12 and 14 months: do the eyes matter? *British Journal of Developmental Psychology* 20: 225-39

Caron AJ, Caron RF, MacLean DJ. 1988. Infant discrimination of naturalistic emotional expressions: the role of face and voice. *Child Dev.* 59: 604-16

Carpenter M, al e. 1998. Fourteen through 18-month-old infants differentially imitate intentional and accidental actions. *Infant Beh. Dev.* 21: 315-30

Carpenter M, al e. 1998. Social cognition, joint attention, and commuicative competence from 9 to 15 months of age. *Monogr. Soc. Res. Child Dev.* 63

Carruthers P, Smith PK, eds. 1996. *Theories of Theories of Mind.* Cambridge: Cambridge University Press. 390 pp.

Casey BJ, Thomas KM, Mccandliss B. 2001 Applications of magnetic resonance imaging to the study of development. In *Handbook of developmental cognitive neuroscience.* Ed. C Nelson, M Luciana. Cambridge, MA: MIT Press

Castelli F, Frith C, Happe F, Frith U. 2002. Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain* 125: 1839-49.

Castelli F, Happe F, Frith U, Frith C. 2000. Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage* 12: 314-25.

Chaminade T, Decety J. 2002. Leader or follower? Involvement of the inferior parietal lobule in agency. *Neuroreport* 13: 1975-8

Chaminade T, Meltzoff AN, Decety J. 2002. Does the end justify the means? A PET exploration of the mechanisms involved in human imitation. *Neuroimage* 15: 318-28

Corbetta M, Kincade JM, Shulman GL. 2002. Neural systems for visual orienting and their relationships to spatial working memory. *J. Cogn. Neurosci.* 14: 508-23

Corkum J, Moore C. 1995. Development of joint visual attention in infants. In *Joint Attention: Its Origin and Role In Development*, ed. C Moore, PJ Dunham, pp. 61-85. Hillsdale, NJ:Erlbaum

Csibra G. 2003. Teleological and referential understanding of action in infancy. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 358: 447-58

Cutting AL, Dunn J. 1999. Theory of mind, emotion understanding, language and familybackground: individual differences and inter-relations. *Child Dev.* 57: 529-48

de Villiers J. 2000. Language and Theory of Mind: what are the developmental relationships? In *Understanding Other minds*, ed. S Baron-Cohen, H Tager-Flusberg, DJ Cohen. Oxford: Oxford University Press

de Villiers J, de Villiers PA. 2000. Linguistic determinism and the understanding of false beliefs. In *Children's Reasoning and the Mind*, ed. P Mitchell, KJ Riggs, pp. 191-228. Hove, England: Psychology Press/Taylor & Francis (UK)

de Villiers J, Pyers JE. 2002. Complements to cognition: a longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive Development* 17: 1037-60

de Villiers JG, de Villiers PA. 2000. Linguistic determinism and false belief. In *Children's Reasoning and the Mind*, ed. P Mitchell, KJ Riggs: Psychology Press

Decety J, Chaminade T. 2003. Neural correlates of feeling sympathy. *Neuropsychologia* 41: 127-38

Decety J, Chaminade T, Grezes J, Meltzoff AN. 2002. A PET exploration of the neural mechanisms involved in reciprocal imitation. *Neuroimage* 15: 265-72.

Decety J, Grezes J. 1999. Neural mechanisms subserving the perception of human actions. *Trends Cogn. Sci.* 3: 172-8

Dehaene-Lambertz G, Dehaene S, Hertz-Pannier L. 2002. Functional neuroimaging of speech perception in infants. *Science*. 298: 2013-5

Dennet D. 1996. *Kinds of Minds: Towards an Understanding of Consciousness*: HarperCollins. 184 pp.

Dennett D. 1978. Beliefs about beliefs. *Behavioural and Brain Sciences* 1: 568-70

di Pellegrino G, Fadiga L, Fogassi L, Gallese V, Rizzolatti G. 1992. Understanding motor events: a neurophysiological study. *Exp. Brain Res.* 91: 176-80

Doherty MJ, Anderson JR. 1999. A new look at gaze: preschool children's understanding of eyedirection. *Cognitive Development* 14: 549-71

Downar J, Crawley AP, Mikulis DJ, Davis KD. 2002. A cortical network sensitive to stimulus salience in a neutral behavioral context across multiple sensory modalities. *J. Neurophysiol.* 87: 615-20

Downing PE, Jiang Y, Shuman M, Kanwisher N. 2001. A cortical area selective for visual processing of the human body. *Science* 293: 2470-3

Dunn J, Brown J. 2001. Emotion, pragmatics and developments in emotion understanding in the preschool year. In *Jerome Bruner: Language, Culture, Self*, ed. D Bakhurst, S Shanker, pp. 88-103. Thousand Oaks, CA: Sage

Farrer C, Franck N, Georgieff N, Frith CD, Decety J, Jeannerod M. 2003. Modulating the experience of agency: a positron emission tomography study. *Neuroimage* 18: 324-33

Farrer C, Frith CD. 2002. Experiencing oneself vs another person as being the

cause of an action: the neural correlates of the experience of agency. *Neuroimage* 15: 596-603.

Feinman S, ed. 1992. *Social Referencing and the Social Construction of Reality in Infancy*. New York, NY: Plenum Press. 424 pp.

Ferstl EC, von Cramon DY. 2002. What does the frontomedian cortex contribute to language processing: coherence or theory of mind? *Neuroimage* 17: 1599-612

Field TM, Walden TA. 1982. Production and perception of facial expressions in infancy and early childhood. *Adv. Child Dev. Behav.* 16: 169-211

Flavell JH. 1988. The development of children's knowledge about the mind: From cognitive connections to mental representations. In *Developing theories of mind.* Ed JW Astington, PL Harris, DR Olson. New York: Cambridge University Press.

Flavell JH. 1999. Cognitive development: children's knowledge about the mind. *Annu. Rev. Psychol.* 50: 21-45

Fletcher PC, Happe F, Frith U, Baker SC, Dolan RJ, et al. 1995. Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension. *Cognition* 57: 109-28.

Frith CD, Frith U. 2000. The physiological basis of theory of mind: functional neuroimaging studies. In *Understanding Other Minds*, ed. S Baron-Cohen, H Tager-Flusberg, DJ Cohen. Oxford: Oxford University Press

Frith U. 2001. Mind blindness and the brain in autism. *Neuron* 32: 969-79

Frith U, Frith CD. 2003. Development and neurophysiology of mentalizing. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 358: 459-73

Gallagher H, Jack A, Roepstorff A, Frith C. 2002. Imaging the intentional stance in a competitive game. *Neuroimage* 16: 814.

Gallagher HL, Frith CD. 2003. Functional imaging of 'theory of mind'. *Trends Cogn. Sci.* 7: 77-83

Gallagher HL, Happe F, Brunswick N, Fletcher PC, Frith U, Frith CD. 2000. Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. *Neuropsychologia* 38: 11-21

Gallese V. 2003. The manifold nature of interpersonal relations: the quest for a common mechanism. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 358: 517-28

Garavan H, Ross TJ, Murphy K, Roche RA, Stein EA. 2002. Dissociable executive functions in the dynamic control of behavior: inhibition, error detection, and correction. *Neuroimage* 17:1820-9

Gergely G, Csibra G. 1997. Teleological reasoning in infancy: the infant's naive theory of rational action. A reply to Premack and Premack. *Cognition* 63: 227-33.

Goel V, Grafman J, Sadato N, Hallett M. 1995. Modeling other minds. *Neuroreport* 6: 1741-6.

Gopnik A, Astington JW. 1988. Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Dev.* 59: 26-37

Grezes J, Decety J. 2001. Functional anatomy of execution, mental simulation, observation, and verb generation of actions: a meta-analysis. *Hum. Brain Mapp.* 12: 1-19.

Hamann S. 2003. Nosing in on the emotional brain. *Nat. Neurosci.* 6: 106-8

Happe F. 1992. Communicative competence and theory of mind in autism: a test of relevance theory. *Cognition* 48: 101-19

Harris PL. 1989. *Children and Emotion: The Development of Psychological Understanding*. Cambridge, MA: Basil Blackwell, Inc. 243 pp.

Harris PL, Johnson C, Hutton D, Andrews G, Cooke T. 1989. Young children's theory of mind and emotion. *Cognition and Emotion* 3: 379 - 400

Haxby JV, Hoffman EA, Gobbini MI. 2000. The distributed human neural system for face perception. *Trends Cogn. Sci.* 4: 223-33

Haxby JV, Hoffman EA, Gobbini MI. 2002. Human neural systems for face recognition and social communication. *Biol. Psychiatry* 51: 59-67

Heberlein AS, Adolphs R, Tranel D, Damasio H. submitted. A dissociaton between emotion recognition and personality attribution from pointlight walkers.

Heiser M, Iacoboni M, Maeda F, Marcus J, Mazziotta JC. 2003. The essential role of Broca's area in imitation. *Eur. J. Neurosci.* 17: 1123-28

Hoffman EA, Haxby JV. 2000. Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nat. Neurosci.* 3: 80-4

Hood BM, Willen JD, Driver J. 1998. Adult's eyes trigger shifts of visual attention in human infants. *Psych. Science* 9: 131-34

Jellema T, Baker CI, Wicker B, Perrett DI. 2000. Neural representation for the perception of the intentionality of actions. *Brain Cogn.* 44: 280-302

Jiang Y, Kanwisher N. IP. Common neural substrates for response selection across modalities and mapping paradigms. *J. Cogn. Neurosci.*

Johnson SC. 2000. The recognition of mentalistic agents in infancy. *Trends Cogn. Sci.* 4: 22-8

Johnson SC. 2003. Detecting agents. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 358: 517-28

Johnson SC, Booth A, O'Hearn K. 2001. Inferring the goals of a non-human agent. *Cognitive Development* 16: 637-56

Johnson SC, Slaughter V, Carey S. 1998. Whose gaze will infants follow? Features that elicit gaze-following in 12-month-olds. *Dev. Sci* 1: 233-38

Just MA, Carpenter PA, Keller TA, Eddy WF, Thulborn KR. 1996. Brain activation modulated by sentence comprehension. *Science* 274: 114-6

Kanwisher N, McDermott J, Chun MM. 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* 17: 4302-11

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: 981-91

Koski L, Wohlschlager A, Bekkering H, Woods RP, Dubeau MC, et al. 2002. Modulation of motor and premotor activity during imitation of target-directed actions. *Cereb. Cortex* 12: 847-55

Lagattuta KH, Wellman HM. 2001. Thinking about the past: early knowledge about links between prior experience, thinking, and emotion. *Child Dev.* 72: 82-102

LeDoux JE. 2000. Emotion circuits in the brain. *Ann. Rev. Neurosci.* 23: 155-84

Legerstee M, Barillas Y. 2003. Sharing attention and pointing to objects at 12 months: is the intentional stance implied? *Cognitive Development* 18: 91-110

Leslie A. 1994. A theory of ToMM, ToBy, and Agency: Core architecture and domain specificity. In *Mapping the Mind: Domain Specificity in Cognition and Culture*, ed. L Hirschfeld, S Gelman, pp. 119-48. New York: Cambridge University Press

Leslie A. 2000. 'Theory of Mind' as a mechanism of selective attention. In *The New Cognitive Neurosciences*, ed. M Gazzaniga, pp. 1235-47. Cambridge, MA: MIT Press

Leslie A, Polizzi P. 1998. Inhibitory processing in the false belief task: two conjectures. *Developmental Science* 1: 247-53

Leslie A, Thaiss L. 1992. Domain specificity in conceptual development. *Cognition* 43: 225-51 Malle BF. 2001. Folk explanations of intentional action. In *Intentions and Intentionality: Foundations of Social Cognition*, ed. BF Malle, LJ Moses, DA Baldwin. Cambridge MA: The MIT Press

Malle BF. IP. The relation between language and theory of mind in development and evolution. In *The Evolution of Language from Pre-language*, ed. T Givon, BF Malle. Amsterdam: Benjamins

Malle BF, Knobe J. 2001. The distinction between desire and intention: A folk-conceptual analysis. In *Intentions and Intentionality: Foundations of Social Cognition*, ed. BF Malle, LJ Moses, DA Baldwin. Cambridge MA: The MIT Press

Malle BF, Moses LJ, Baldwin DA, eds. 2001. *Intentions and Intentionality: Foundations of Social Cognition*. Cambridge MA: The MIT Press

McCabe K, Houser D, Ryan L, Smith V, Trouard T. 2001. A functional imaging study of cooperation in two-person reciprocal exchange. *Proc. Natl. Acad. Sci. U. S. A.* 98: 11832- 5.

McCarthy G, Puce A, Gore JC, Allison T. 1997. Face-specific processing in the human fusiform gyrus. *J. Cogn. Neurosci.* 9: 605-10

Meltzoff AN. 1995. Understanding the intentions of others: re-enactment of intended acts by 18- month-old children. *Dev. Psychol.* 31: 838-50

Meltzoff AN, Decety J. 2003. What imitation tells us about social cognition: a rapprochement between developmental psychology and cognitive neuroscience. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 358: 491-500

Milham MP, Banich MT, Webb A, Barad V, Cohen NJ, et al. 2001. The relative involvement of anterior cingulate and prefrontal cortex in attentional control depends on nature of conflict. *Brain Res. Cogn. Brain Res.* 12: 467-73

Moll J, de Oliveira-Souza R, Eslinger PJ, Bramati IE, Mourao-Miranda J, et al. 2002. The neural correlates of moral sensitivity: a functional magnetic resonance imaging investigation of basic and moral emotions. *J. Neurosci.* 22: 2730-6.

Morris JS. 2002. How do you feel? *Trends Cogn Sci* 6: 317-9

Moses LJ. 2001. Executive accounts of theory-of-mind development. *Child Dev.* 72: 688-90.

Moses LJ, Baldwin DA, Rosicky JG, Tidball G. 2001. Evidence for referential understanding in the emotions domain at twelve and eighteen months. *Child Dev.* 72: 718-35

Narumoto J, Okada T, Sadato N, Fukui K, Yonekura Y. 2001. Attention to emotion modulates fMRI activity in human right superior temporal sulcus. *Brain Res. Cogn. Brain Res.* 12: 225-31

Nelson CA. 1987. The recognition of facial expressions in the first two years of life: mechanisms of development. *Child Dev.* 58: 889-909

Nelson K. 1996. *Language in Cognitive Development: Emergence of the Mediated Mind.* New York, NY: Cambridge University Press. 432 pp.

Nichols S, Stich S, Leslie A, Klein D. 1996. Varieties of off-line simulation. In *Theories of Theories of Mind*, ed. P Carruthers, PK Smith. Cambridge: Cambridge University Press

Pellicano E, Rhodes G. 2003. The role of eye-gaze in understanding other minds. *British Journal of Developmental Psychology* 21: 33-43

Pelphrey KA, Singerman JD, Allison T, McCarthy G. 2003. Brain activation evoked by perception of gaze shifts: the influence of context. *Neuropsychologia* 41: 156-70

Perner J. 1991 *Understanding the representational mind.* Cambridge, MA: MIT Press

Perner J, Lang B. 2000. Theory of Mind and executive function: is there a developmental relationship? In *Understanding Other Minds*, ed. S Baron-Cohen, H Tager-Flusberg, DJ Cohen. Oxford: Oxford University Press

Perner J, Stummer S, Lang B. 1999. Executive functions and theory of mind: cognitive complexity or functional dependence? In *Developing Theories of Intention: Social Understanding and Self-control.*, ed. PD Zelazo, JW Astington, e al, pp. 133-52. Mahwah, NJ: Lawrence Erlbaum Associates

Peterson CC, Siegal M. 1995. Deafness, conversation and theory of mind. *J Child Psychol. Psychiat.* 36: 459-74

Phillips AT, Wellman HM, Spelke ES. 2002. Infants' ability to connect gaze and emotional expression to intentional action. *Cognition* 85: 53-78

Povinelli DJ. 2001. On the possibilities of detecting intentions prior to understanding them. In *Intentions and Intentionality: Foundations of Social Cognition*, ed. BF Malle, LJ Moses, DA Baldwin, pp. 444. Cambridge, MA: MIT Press

Premack D, Woodruff G. 1978. Does the chimpanzee have a theory of mind? *Behavioural and Brain Sciences* 1: 515-26

Preston SD, de Waal FBM. 2002. Empathy: Its ultimate and proximate bases. *Behavioural and Brain Sciences* 25: 1-72

Puce A, Allison T, Bentin S, Gore JC, McCarthy G. 1998. Temporal cortex activation in humans viewing eye and mouth movements. *J. Neurosci.* 18: 2188-99

Puce A, Perrett D. 2003. Electrophysiology and brain imaging of biological motion. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 358: 435-45

Repacholi BM, Gopnik A. 1997. Early reasoning about desires: evidence from 14- and 18-montholds. *Dev. Psychol.* 33: 12-21

Rieffe C, Terwogt MM, Koops W, Stegge H, Oomen A. 2001. Preschooler's appreciation of uncommon desires and subsequent emotions. *British Journal of Developmental Psychology* 19: 259-74

Rizzolatti G, Fogassi L, Gallese V. 2001. Neurophysiological mechanisms underlying the understanding and imitation of action. *Nat. Rev. Neurosci.* 2: 661-70.

Roder B, Stock O, Neville H, Bien S, Rosler F. 2002. Brain activation modulated by the comprehension of normal and pseudo-word sentences of different processing demands: a functional magnetic resonance imaging study. *Neuroimage* 15: 1003-14

Ruby P, Decety J. 2001. Effect of subjective perspective taking during simulation of action: a PET investigation of agency. *Nat. Neurosci.* 4: 546-50.

Ruffman T, Slade L, Rowlandson K, Rumsey C, Garnham A. 2003. How language relates to belief, desire and emotion understanding. *Cognitive Development* 113: 1-20

Saxe R, Kanwisher N. IP. People thinking about thinking people: fMRI investigations of theory of mind. *Neuroimage*

Saxe R, Xiao DK, Kovacs G, Perrett D, Kanwisher N. submitted. Distinct representations of bodies, actions and thoughts in posterior superior temporal sulcus.

Schultz RT, Grelotti DJ, Klin A, Kleinman J, Van der Gaag C, et al. 2003. The role of the fusiform face area in social cognition: implications for the pathobiology of autism. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 358: 415-27

Searle J. 183. *Intentionality: An Essay in the Philophy of Mind.* New York: Cambridge University Press

Shulman GL, d'Avossa G, Tansy AP, Corbetta M. 2002. Two attentional processes in the parietal lobe. *Cereb. Cortex* 12: 1124-31

Siegal M, Varley R. 2002. Neural systems involved in "theory of mind". *Nat. Rev. Neurosci.* 3: 463-71

Stich S, Nichols S. 1998. Theory theory to the max. *Mind & Language* 13: 421-49

Stroop JR. 1938. Factors affecting speed in serial verbal reactions. *Psychol. Monogr.* 50: 38-48

Sylvester CY, Wager TD, Lacey SC, Hernandez L, Nichols TE, et al. 2003. Switching attention and resolving interference: fMRI measures of executive functions. *Neuropsychologia* 41: 357-70

Tager-Flusberg H, Sullivan K. 2000. A componential view of theory of mind: evidence from Williams syndrome. *Cognition* 76: 59-90.

Terwogt MM, Stegge H. 1998. Children's perspective on the emotion process. In *The Social Child.*, ed. A Campbell, S Muncer, pp. 249-69. Hove, England: Psychology Press/ Erlbaum (UK)

Tomasello M, Strosberg R, Akhtar N. 1996. Eighteen-month-old children learn words in nonostensive contexts. *J. Child Lang.* 23: 157-76

Tootell RB, Reppas JB, Kwong KK, Malach R, Born RT, et al. 1995. Functional analysis of human MT and related visual cortical areas using magnetic resonance imaging. *J. Neurosci.* 15: 3215-30

Tremoulet PD, Feldman J. 2000. Perception of animacy from the motion of a single dot. *Perception* 29: 943-51

Vandenberghe R, Nobre AC, Price CJ. 2002. The response of left temporal cortex to sentences. *J. Cogn. Neurosci.* 14: 550-60

Varley R, Siegal M. 2000. Evidence for cognition without grammar from causal reasoning and 'theory of mind' in an agrammatic aphasic patient. *Curr. Biol.* 10: 723-6.

Varley R, Siegal M, Want SC. 2001. Severe impairment in grammar does not preclude theory of mind. *Neurocase* 7: 489-93

Vogeley K, Bussfeld P, Newen A, Herrmann S, Happe F, et al. 2001. Mind reading: neural mechanisms of theory of mind and self-perspective. *Neuroimage* 14: 170-81.

Watson AC, Painter KM, Bornstein MH. 2001. Longitudinal relations between 2-year-olds' language and 4-year-olds' theory of mind. *J. Cog. and Dev.* 2: 449-57

Wellman HM, Cross D. 2001. Theory of mind and conceptual change. *Child Dev.* 72: 702-7.

Wellman HM, Cross D, Watson J. 2001. Meta-analysis of theory-of-mind development: the truth about false belief. *Child Dev.* 72: 655-84.

Wellman HM, Harris PL, Banerjee M, Sinclair A. 1995. Early understanding of emotion: evidence from natural language. *Cognition and Emotion* 9: 117-49

Wellman HM, Lagattuta KH. 2000. Developing understandings of mind. In *Understanding Other Minds*, ed. S Baron-Cohen, H Tager-Flusberg, DJ Cohen

Wellman HM, Woolley JD. 1990. From simple desires to ordinary beliefs: the early development of everyday psychology. *Cognition* 35: 245-75

Whalen PJ. 1999. Fear, vigilance and ambiguity: initial neuroimaging studies of the human amygdala. *Curr. Directions Psych. Sci.* 7: 177-87

Wicker B, Michel F, Henaff MA, Decety J. 1998. Brain regions involved in the perception of gaze: a PET study. *Neuroimage* 8: 221-7

Wicker B, Perrett DI, Baron-Cohen S, Decety J. 2003. Being the target of another's emotion: a PET study. *Neuropsychologia* 41: 139-46

Wimmer H, Perner J. 1983. Beliefs about beliefs: representation and constraining

function of wrong beliefs in young children's understanding of deception. *Cognition* 13: 103-28

Winston JS, Strange BA, O'Doherty J, Dolan RJ. 2002. Automatic and intentional brain responses during evaluation of trustworthiness of faces. *Nat. Neurosci.* 5: 277-83.

Wolpert DM, Doya K, Kawato M. 2003. A unifying computational framework for motor control and social interaction. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* 358: 593-602

Woodward AL. 1998. Infants selectively encode the goal object of an actor's reach. *Cognition* 69: 1-34

Woolfe T, Want SC, Siegal M. 2002. Signposts to development: Theory of mind in deaf children. *Child Dev.* 73: 718-78

Zacks JM, Braver TS, Sheridan MA, Donaldson DI, Snyder AZ, et al. 2001. Human brain activity time-locked to perceptual event boundaries. *Nat. Neurosci.* 4: 651-5

Zaitchik D. 1990. When representations conflict with reality: the preschooler's problem with false beliefs and "false" photographs. *Cognition* 35: 41-68

Zaitchik D. 1991. Is only seeing really believing? Sources of the true belief in the false belief task. *Cognitive Development* 6: 91-103

Chapter 4. What is a theory of mind?

1 Introduction

I have so far left untouched the most fundamental questions about theory of mind. Consider psychological concepts, like 'goal', 'belief', or 'repressed desire', or psychological 'laws', like 'seeing leads to knowing', 'people act to achieve their goals given their beliefs', or 'people sometimes believe what they want to believe'. Do we have, and use, a mental representation of these concepts (or 'laws') in our everyday reasoning about other people? If so, where they come from? What determines their content? How does that content change?

In the domain of theory of mind, there are three basic conceptions of mental structure that propose to answer these questions: modules, theories and off-line simulation. Each of these three frameworks offers an account of the origin, the implementation, and the developmental trajectory of reasoning about other minds. I will not have room in this chapter to survey all of the empirical and theoretical arguments that have been mustered for and against (particular versions of) each proposal. Instead, I will take a narrower perspective, and consider the extent to which the claims of these theoretical models can be tested using functional magnetic resonance imaging (fMRI).

This chapter is divided into two main sections. In the first section I contrast modules with theories, and test these two frameworks against the fMRI results in the first three chapters of this dissertation. My conclusion weakly favours a model combining both a module and a theory: Carey and Spelke's (1994) Core Knowledge hypothesis. The first section thus provides an illustration of the potential theoretical impact of fMRI. The second section offers a cautionary tale, though. Proponents of the third framework, off-line simulation (e.g. Gallese and Goldman 1998), have claimed support from fMRI results. However, I will argue that these claims are premature at best, and indeed that the off-line simulation proposal may be too under-specified as it stands for a test by fMRI to even be possible. Consequently, in the Appendix, I have included stronger theoretical arguments to show that off-line simulation cannot be the whole story about the theory of mind.

2 Modules and Theories

Both modules and theories share a commitment to a mental structure based on "systems of rules and representations" (Chomsky 1980). On each of these models, humans have representations of psychological concepts and 'laws', including those listed above, and use these mental representations, for instance when trying to explain an

observed behaviour, like why Sarah wrote her boss a $20 cheque. Modules and theories typically differ, though, in the proposed origin of these concepts, and the mechanisms for conceptual change. Here, I will provide a rough sketch of modules and theories, focussing on those differences between the two frameworks that are most likely to be relevant for testing with fMRI.

2.1 Modules

The classic account of a modular architecture was described by Fodor (1983). The essential feature of a module is restricted information flow: there must be some information outside the module, to which operations within the module do not have access, and/or information within the module to which other, external cognitive operations do not have access (cf. Scholl and Leslie 1999). Consonant with this informational limitation, modules are "domain-specific" – that is, they operate on only certain kinds of inputs. Modules may often be innate (in some sense) and depend on distinct and specialised neural real estate. Other 'symptoms' of modularity include fast and mandatory processing, and relatively 'shallow' outputs.

One popular criticism of modularity is that it is somehow non- or even anti-developmental, unable to account for changes in the child's knowledge within the domain of the module. Thus the assumption has been that a modularity thesis must explain transitions in task performance in terms of factors strictly external to the module. Scholl and Leslie (1999) show persuasively that this is a misconception, and describe at least three different mechanisms for development of a modular system: (1) modules may come 'on-line' when triggered by the environment, even though the essential character of the module is fixed. Thus a modular system may be made up of multiple modules, each maturing independently. (2) The character of a module may be simultaneously determined by the environment and highly constrained by its own innate structure. An example of this is the "parameter setting" thought to account for the development of the syntactic module in different linguistic contexts. (3) The character of a module may have an innate basis which is shaped by module-internal development. In each of these cases, "these developments are determined by the genetic endowment, though the precise manner in which the genetic plan is realised depends in part on external factors" (Chomsky 1980).

Leslie (1994) has proposed that all of the fundamental concepts of theory of mind originate in a Theory of Mind Module (ToMM), including concepts of 'goal', 'agent', 'perception', 'pretence', 'desire', 'belief'. The ToMM "spontaneously and post-perceptually attends to behaviours and infers (i.e. computes) the mental states that contributed to them" (Scholl and Leslie 1999). Within this framework developmental

changes in theory of mind performance occur through at least two separate mechanisms[13]. First, Leslie (1994) proposed that theory of mind relies on two distinct theory of mind modules. ToMM1, which matures earliest, is responsible for the "actional" properties of other people: those that let them act in pursuit of goals and react to the environment, presumably including concepts of 'perception', 'agent' and 'goal'. ToMM2, which develops soon after the first birthday, represents "cognitive" properties like 'pretence', 'belief' and 'desire'. To explain the delay between the maturation of ToMM2 (around age one) and the first success on the false belief task (around age three and a half), Leslie (e.g. Roth and Leslie 1998, Leslie 2000) appeals to an independent module, the selection processor (SP). The SP helps children to resist default assumptions, such as the assumption that beliefs are true; however the SP does not fully mature until approximately the child's fourth year.

2.2 Theories

The dominant alternative mental structure proposed to explain the origin and use of concepts is quasi-scientific explanatory theory. Theories and modules differ most clearly in the proposed mechanism (and process) of conceptual change[14]. Modules are usually envisioned as developing "along an internally directed course under the triggering and partially shaping effect of the environment" (Chomsky 1980). Thus the environment may determine the timetable of development, and in some cases (as in linguistic parameter setting) some of the particular structure, of a module. However, informational limitations and the genetic endowment ensure that external influence is relatively narrow and constrained, keeping the modular structure consistent across individuals. It is therefore one of the strengths of modularity that it can naturally account for the universality of the concepts that children develop (Scholl and Leslie 1999). By contrast, "the most important thing about theories is [...] their defeasibility. Theories may turn out to be inconsistent with the evidence, and because of this theories change" (Gopnik and Meltzoff 1997). Gopnik and Meltzoff (1997) provide a sketch of how one theory succeeds another: children (and grown-ups) add auxiliary hypotheses to the original

---

[13] Scholl and Leslie (1999) provide a convincing argument against the development of theory of mind by parameter-setting.

[14] This discussion will necessarily be a gross over-simplification of all of the theoretical positions held by self-styled proponents of modules and theories. Also, I discuss none of Gopnik and Meltzoff's (1997) "structural" and "functional" characteristics of theories here. These include appeal to unobserved entities, ontological commitment, and a coherent causal explanatory structure. However, all of these structural and functional features could, in principle, be instantiated in a module. For this reason, Gopnik and Meltzoff (1997) and I focus on the "dynamic features" of theories, the mechanisms for theory change.

theory until it becomes too unwieldy and/or a new alternative becomes available, at which point the two competing theories can be tested through intense observation and experimentation. Theoretical revision is therefore susceptible to very broad influence of the environment; theory change may transform the child's knowledge of a domain beyond recognition, producing incommensurability between earlier and later concepts (Carey 1985).

The distinction between theories and modules, thus conceived, is not sharp but it may be useful. For instance, I think that Gopnik and Meltzoff's account of theory change, and Leslie's modularity, make different predictions about the fate of an out-grown concept. Modules stick around, so even if a concept is inaccurate or unsatisfactory, it will continue to be applied by the module whenever the module receives appropriate input. Changes are effected mostly by adding new modules, or by changing the interpretation of the module's output (Scholl and Leslie 1999), not by destruction or modification of the module itself. Theories, on the other hand, can be altered directly. Unsatisfactory concepts or theories are replaced or changed. Gopnik and Meltzoff (1997) make this feature of theory development explicit: "We propose that there are innate theories that are later modified and revised. The process of theory change and replacement begins at birth. To continue [the] metaphor [of knowledge as a boat under repair throughout the voyage], innate theories are the boats that push off from the pier. The boat you start out with may have a considerable effect on the boat you end up with, even if no trace of the original remains" (Gopnik and Meltzoff 1997, p. 51).

Modules and theories are not mutually exclusive. Leslie himself claims "only that ToM has a specific innate *basis*" (Scholl and Leslie 1999, original emphasis). During development, children may use their general capacities for causal learning and hypothesis testing to become increasingly sophisticated in their interpretation and application of the *outputs* of the ToM modules. Carey and Spelke (1994) allow an even larger contribution to theory and theory change. According to their Core Knowledge hypothesis, a few domain-specific modules provide an innate conceptual endowment to get children started on conceptual representations of the world, but children must develop theories to complement, extend and integrate the relatively shallow outputs of the modules. In the domain of theory of mind, for instance, an innate module is responsible for the early-developing concepts 'agent', 'perception' and 'goal', not unlike Leslie's ToMM1 (1994). However, later-developing concepts including, paradigmatically, 'belief' are not provided by the module and must be developed in the context of an intuitive theory of psychology.

Gopnik (1996) defends a particularly extreme position on the role of theories, and theory change, in the development of theory of mind: namely, that it is "theories all the way down" (see Stich and Nichols 1998, for a very good discussion of this position).

"Infants seem to have innate knowledge ... and this knowledge is theory-like," Gopnik (1996) claims, and so the process of hypothesis testing sketched above is responsible for all the progress in knowledge acquisition that infants make after birth. Concepts like 'agent', 'perception' and 'goal' are included in an early (perhaps innate) version of the infant's "theory of action"; this same theory is later revised to include concepts like 'belief' when that revision is necessary to accommodate the results of the child's observation and experimentation.

2.3 Modules and theories and fMRI

Modules and theories are easiest to distinguish by the extent of external influence on the process of conceptual change. FMRI provides access to a static and very coarse description of the structure of a cognitive function in the brains of adults. Is it possible to bridge these two levels of description, to allow meaningful tests of the theoretical alternatives using fMRI? In order to do so, we need bridging laws, essentially translations from the theoretical commitments of one level of description to predictions for a different level. For the purposes of this discussion, I therefore propose a bridging law that (while vague and disputable) seems like a plausible first pass at the intuitions behind both modules and theories:

> (BL): A given domain-specific module (or theory) will be implemented in the brain by either (a) a single, specially dedicated brain region, or (b) a set of specially dedicated brain regions, or (c) widely distributed neural mechanisms, not clustered into regions.

"Specially dedicated", in both (a) and (b), imply that the same brain region(s) is (/are) not shared by other modules or theories. Dedicated brain regions are more often associated with modules, and distributed (unclustered) mechanisms may seem like an intuitive implementation of theories, but these associations need not hold. Proponents of modularity have explicitly endorsed the idea that a module may be implemented by distributed neurones (e.g. Scholl and Leslie 1999). And, since the mechanisms and the rate by which functions come to be implemented in spatially contiguous brain regions are unknown, a theory might conceivably colonise a specific, restricted piece of neural real estate for any number of reasons.

Weak though it is, BL is sufficient to let us turn the theoretical differences between modules and theories described above into a prediction testable by fMRI. As described above, modules stick around. Leslie's (1995) division of theory of mind concepts between two modules, ToMM1 and ToMM2, therefore predicts that if any dedicated regions for theory of mind exist in the brain (i.e. option C above does not hold), then there will be at least two distinct brain regions, or sets of regions, involved in theory

of mind: one to implement ToMM1 and another to implement ToMM2. This prediction is exactly consistent with the results described in the previous chapters. In the brains of adult humans, one set of brain regions, including regions of the temporo-parietal junction and medial frontal cortex, seems to be involved in the attribution of "cognitive properties" like beliefs – the domain of ToMM2. Different brain regions, including regions of posterior superior temporal sulcus and lateral inferior frontal cortex, are associated with representations of perception and goals.

Carey and Spelke's (1994) core knowledge hypothesis also predicts my results, although in that case the theoretical division lies between a module (for the earlier-developing concepts like 'goal') and a theory for the rest (including 'belief'). In fact, the core knowledge hypothesis may have a slight edge over Leslie's particular modular architecture, with respect to the imaging results. Leslie (1995, 2000, Roth and Leslie 1998) proposes that a third module, the Selection Processor (SP), is required for understanding both false beliefs and "false" (i.e. out-of-date) photographs, but not for the attribution of true beliefs. It is the slow maturation of the SP that explains the failure on the false belief task of young children who already have a ToMM2. However, I did not find any brain region with this predicted functional profile. Every brain region involved in the attribution of false beliefs also showed a high response during the attribution of true beliefs (chapters 1 and 3). The core knowledge hypothesis predicts only two components of theory of mind, and so is broadly consistent with all aspects of my data.

Gopnik's (1996) more extreme position fares less well under the same assumptions. Gopnik proposes that infants initially have an immature "theory of action," which is revised and changed during development, possibly until "no trace of the original remains" (Gopnik and Meltzoff 1997). This proposal, combined with BL, predicts that theory of mind is implemented in the brain either by a single brain region, or by a set of brain regions, or by a distributed set of neural mechanisms. However, it does not predict that theory of mind is implemented by *two* sets of dedicated brain regions, as I have shown.

This is not, of course, a knock-down argument against the "theories all the way down" position. Gopnik and Meltzoff (1997) do not themselves propose any bridging law between mental and neural entities, so they may simply deny BL. Alternatively, the "theory of action" could be split into two separate theories, perhaps a theory of action and a theory of mind. Or finally, a proponent of theories could allow that later theories exist alongside and in addition their predecessors, rather than modifying or replacing the earlier theories (e.g. Perner 1991). Nor is Leslie's (e.g. 2000) Selection Processor proposal necessarily untenable because we did not observe it through fMRI. Adults may have become so familiar with false beliefs that the contribution of the SP is no longer

necessary. What's more, negative results must be interpreted only very cautiously; there are many ways to miss something with fMRI.

In all, fMRI cannot yet provide a definitive test of theoretical frameworks for cognition. I hope rather that these arguments will play two relatively modest roles: a constraining role and a motivating role. Constraining because, insofar as BL accurately captures the intention of the module- and theory- based frameworks, future models of the mental structure of theory of mind must incorporate an explanation of the two parts of its neural implementation. (For instance, Gopnik may have to adjust her picture of a single 'theory of action' operating "all the way down.") And motivating because I hope that this illustration will encourage the theoreticians to make explicit the neural predictions of their frameworks and the cognitive neuroscientists to test them. The contribution of fMRI must be severely limited as long as the theoretical description of the processes under investigation is under-specified. In the next part of this chapter, I will argue that just such under-specification is at fault for the inadequacy of fMRI to test the claims of the third proposed framework for theory of mind: off-line simulation.

3 Off-line simulation

The third framework for theory of mind, off-line simulation, challenges the assumption common to both modules and theories, that a theory of mind depends on the representation and use of psychological concepts and 'laws'. The central feature of the off-line simulation account of theory of mind is that the observer uses her own mind as an analogue model of the mind of the actor. If the causal structures of the two minds are relevantly similar, then the (hypothetical, or 'pretend') response of the observer's mind to a set of input conditions will accurately predict the (actual) response of the actor's mind to the same conditions.

Critically, a simulator thus uses information about *how* minds function, without explicitly representing *that* minds function in that way (Perner 1998). Gallese and Goldman (1998) describe the proposed "cognitive steps in predicting or explaining someone's decision by means of simulation" as follows: "A dedicated pretend-state generator generates pretend beliefs and desires suited to the target agent. These pretend beliefs and desires are fed into the attributor's decision-making system – the same system that normally operates on natural non-pretend beliefs and desires. The output of the decision-making system is taken 'off-line'. That is, instead of being fed into the action control system, the output decision is sent to the behaviour-predicting and -explaining system, which outputs a prediction that the target will make that very decision."

The version of the simulation framework that is most clearly different from modules and theories is called 'off-line' simulation, by contrast with 'information-based'

simulation (Nichols et al 1998; Goldman 1992 makes a similar contrast between "process-driven" and "theory-driven" simulations). In building an information-based simulation, a computational modeller must have explicit access to the (putative) causal structure of the system whose behaviour she wishes to predict. That is, the modeller must have a *theory* of the modelled system. An off-line simulator, by contrast, simply possesses (for other reasons) a functioning model– in this case, the simulator's own mind – and uses it. There is no need to represent, for instance, anything like a principle of rationality (that a person will generally act to achieve his desires given his beliefs). When particular beliefs and desires are provided as 'pretend' inputs, the local decision making system will provide the rational action choice as its output. Thus, an off-line simulator does not need or use any theoretical knowledge about the mind.

The off-line simulation model obviously cannot countenance the kind of conceptual development in the theory of mind, 'from cognitive connections to mental representations' (Flavell 1988) endorsed in chapter 3. Instead proponents of off-line simulation explain the developmental trajectory, from initial concepts of perceptions and goals to the later more sophisticated understanding of false belief, in terms of the competence of the simulator and the extraneous demands of the simulation. Simulation occurs against a background of the child's actual beliefs and desires, called the default setting. The number of default settings that require concurrent temporary suspension determines the difficulty of a given simulation. Children's competence and flexibility with suspending their own beliefs improve with age, and their performance on the false belief task follows suit (Harris 1989, 1992).

The intuitive appeal of off-line simulation lies in its (supposed) cognitive economy, and in its power to explain three features of theory of mind: the vividness and immediacy of a shared experience, the inability of practitioners or proponents of theory of mind to articulate the principles and rules of their theory, and the actual mechanics of everyday behaviour predictions (these claims are discussed in depth in the Appendix). However, its proponents also argue that off-line simulation is strongly supported by the recent neuroscientific discovery of mirror neurones: cells (in macaques, or brain regions in humans) that are active during both observation and execution of the same action (e.g. Gallese and Goldman 1998). It is to these claims about neuroscientific evidence that I turn now.

3.1 Mirror neurones and off-line simulation

The primary thesis of off-line simulation is that action explanation and action planning - inferences about others' decisions and making one's own - rely on the same mechanisms. The form of this prediction is particularly well suited to resolution by

neuroimaging techniques, as described in chapter 3, section 2, and proponents of off-line simulation have claimed some success. A growing body of neuroimaging work provides clear evidence that the same neural and functional mechanisms subserve action observation and execution.

The recent human neuroimaging research was sparked by the discovery of "mirror neurones" in macaque monkeys. The premotor cortex of macaque monkeys contains neurones that code whole action sequences (by contrast to primary motor cortex neurones, which code simpler motor primitives). For instance, a class of neurones in areas F5 and F6 fire when the monkey reaches to grasp an object; stimulating the same neurones produces a coherent reaching-grasping sequence (Rizzolatti et al 1990).The critical new discovery is that many of these so-called 'motor' neurones also have visual response properties. These "mirror-neurones" fire equally when the monkey executes a particular action, and when the monkey observes someone else executing the same action (di Pellegrino et al 1992). Since their first discovery in premotor cortex, mirror neurones have also been reported in anterior intraparietal cortex (e.g. Murata et al 1996).

Many mirror neurones are highly selective, firing only to a particular kind of movement or grasping action; typically the neurone is selective for the *same* kind of action whether executed by the monkey or by the experimenter (e.g. Rizzolatti et al 2000). Most impressively, mirror neurones seem to fire in response not to a pattern of motion, but to an abstract representation of a goal directed action. Umilta et al (2001) showed monkeys a reaching action, either into empty space, or towards an apple. Critically, in half of the trials, the target location was occluded before the reaching occurred – so that the reaching towards the (occluded) apple or empty space were visually identical at the time of the motion. Some mirror neurones fired only when the monkey knew that there was an apple present. That is, they responded only when the observed (visually identical) motion could be interpreted as goal-directed. In all cases, these same neurones fired when the monkey made a goal-directed reaching action himself. Finally, some mirror neurones fire to a goal-directed action even when the entire action is invisible. Kohler et al (2002) showed that mirror neurones fire to the characteristic sound of an action, as well as to execution or observation of the same action; selectivity in all three modalities was congruent.

The studies described above provide strong evidence that in monkeys, action execution and observation produce activity in the same neurones. In order to count in favour of off-line simulation in human reasoning, the same must be true in the human brain. And indeed, neuroimaging studies have begun to produce evidence that classic 'motor' regions in human cortex, as in monkeys, also have visual response properties. Observing the motor actions of another person produced increased blood flow in human

dorsal premotor regions (Buccino et al 2001, Chaminade et al 2002, Decety et al 1997, Grezes and Decety 2001, Rizzolatti et al 1996, Iacoboni et al 2001).

Further confirmation comes from studies using TMS (transcranial magnetic stimulation). Strafella et al (2000) used TMS to stimulate primary motor cortex, and recorded the motor evoked potentials (MEP) at two muscular recording sites – in the biceps and FDI (first dorsal interosseous) muscles – while subjects were at rest, or watching the experimenter model hand and arm motions. Strafella et al found that action observation modulated the TMS-elicited activity in the muscles, suggesting that action observation modulates excitability in primary motor cortex. Furthermore, the increased excitability was specific: excitation increased selectively in the muscles congruent with the actions observed, in the biceps for arm motion and in the FDI for fine hand motion. Gangitano et al (2001) found that the amplitude of the TMS induced MEP at the FDI was modulated by the size of the observed finger aperture.

Thus neuroimaging results provide converging evidence that in humans, as in monkeys, motor cortex is active significantly and selectively during observation of actions, and that the visual selectivity is congruent with the motor selectivity of the same regions. But are these common representations actually necessary for comprehension of observed actions and executing one's own actions, or might one of these functions activate these brain regions merely epiphenomenally? There is some behavioural evidence against epiphenomenal co-activation. When an observed action is incongruent with the action that must be simultaneously executed, execution is seriously impaired. The interference produced by incompatible gestures is much larger than that produced by other kinds of distractors (Brass et al 2001), even when the hand gestures are task-irrelevant and the other distractors are potentially task-relevant (Sturmer et al 2000). This interference suggests that the representation of an observed action (by someone else) is competing for the same resources that are necessary for representing one's own to-be-executed action.

In all, a strong case has been assembled from both behavioural and neuroscientific studies in support of common neural mechanisms for the execution and perceptual representation of goal-directed manual actions, consistent with the predictions of off-line simulation. Does this evidence establish off-line simulation as the correct framework for human theory of mind? I think not.

Even the relatively limited claim that action comprehension and action execution rely on common brain mechanisms may be in some empirical trouble. Ideomotor apraxia, a neuropsychological condition associated with brain damage in motor and parietal cortices, is characterised by severe deficits in action execution, in the absence of

paralysis or inability. In particular ideomotor apraxics have difficulty in demonstrating goal-directed actions, with tools or in mime, to verbal instructions, and in imitation. However, studies of ideomotor apraxia suggest that action execution and action recognition can be neurally dissociated from one another. In both individual case studies, and across groups of subjects, a severe deficit in action execution coexists with a relatively unimpaired ability to recognise, comprehend and discriminate the same actions when modelled by someone else (Buxbaum et al 2001, Halsband et al 2001, Hanna-Pladdy et al 2001, Karekenen et al 1998, Toraldo et al 2000). That is, action interpretation can survive a deficit of action execution, contrary to the asymmetric dependence proposed by off-line simulation.

However, there are deeper problems for the purported relationship between off-line simulation and mirror mechanisms in the brain. First, the very homology between monkeys and humans that has driven the discovery of mirror neurone mechanisms renders these mechanisms suspicious as an account of theory of mind, because macaque monkeys are very poor candidates for the possession of a theory of mind (e.g. Povinelli and Giambrone 2001).

The second, and related, problem is that the central task for a theory of mind is to predict and explain human behaviour. In many cases, this relies on the ability to identify the *reasons* for an action (Malle 1999), based on mental states like beliefs and desires. While mirror neurones may be able to represent the (physically present) target of an action, there is certainly no hint of a mirror mechanism that could represent the reason for the action. This is, of course, particularly clear when an action proceeds based on a false belief. It is an empirical question, but I doubt that a mirror neurone would respond to a reach towards empty space, if the actor *thought* that there was an apple present.

As I understand it, the argument based on mirror neurones that off-line simulation is the right framework for theory of mind must go something like this: since executing and understanding goal-directed manual actions rely on a common mechanism, perhaps executing and understanding thoughts will do the same.[15] But this is weak support indeed.

---

[15] A different argument, sometimes advocated by the discoverers of mirror neurones, is that the mirror neurone system provides a mechanism for the direct perception of the mental causes of behaviour. When confronted by another person's action, our own motor planning and execution system "resonates" with the observed action (Gibson 1966, Rizzolatti et al 2001). This move, like Gibson's original proposal, is meant to use the information available in the "array" to finesse the need for "information processing" – in this case, for reasoning about other minds. But I think resonance is as explanatorily empty in the domain of understanding other minds as in the domain of detecting image invariants. Marr's (1982) reply to Gibson applies equally well here: explaining an observed behaviour is "exactly and precisely an information

To take a step back, then: if the human mirror system does not constitute neuroscientific evidence for off-line simulation as the mechanism for theory of mind, are there other neuroimaging results that either do, or could, provide such evidence? After all, as I stated earlier, the formulation of the off-line simulation model seems to be very well-suited to testing by neuroscientific means. Nevertheless, I am inclined to answer that there are not[16]. For a claim that two processes rely on a single neural mechanism to be evaluated by neuroimaging, there must be a way to independently identify the brain regions associated with each individual process. The off-line simulation claim that "explaining and predicting behaviour" relies on the same mechanisms as one's own "decision making system" fails this test on both counts.

The first difficulty is that "explaining and predicting behaviour" relies on multiple distinct processes, and the off-line simulation proposal provides no way to distinguish which of these functions is responsible for a given observed brain activation. For instance, according to the most clearly specified version of off-line simulation, endorsed by Nichols et al (1998) and Gallese and Goldman (1998), the simulation procedure begins as follows: "A dedicated pretend-state generator generates pretend beliefs and desires suited to the target agent. These pretend beliefs and desires are fed into the attributor's decision-making system." Subsequently "the output decision is sent to the behaviour-predicting and -explaining system" (Gallese and Goldman 1998). However, these authors suggest no way to determine whether a particular brain region, observed to be highly active during a "theory of mind" task, reflects the function of the "pretend-state generator," the "decision-making system," or the "behaviour-predicting and -explaining system."

---

processing problem," and to think that it can be explained by an appeal to resonance is to "vastly underrated the sheer difficulty" of it.

[16] One paper that might seem to address this issue is Vogeley et al (2001). These authors have subjects read short vignettes about human actions – the same vignettes used in other neuorimaging studies of theory of mind (Fletcher et al 1995, Gallagher et al 2000) – written either in the third person ('he', 'John') or in the second person ('you'). Brain regions associated with belief attribution in the standard third person perceptive, including in particular the temporo-parietal junction – were also strongly activated by stories in the second person. These results may seem to suggest that reasoning about others relies on the same brain regions as reasoning about one's self. However, this is not sufficient to support the off-line simulation claim. Developmental psychologists have long observed that one's own past or hypothetical beliefs are just as hard to predict and explain as the beliefs of others; comprehension of one's own and other's false beliefs follow precisely the same developmental timetable (e.g. Gopnik and Astington 1988). The distinctive claim of off-line simulation is that understanding mental states and reasoning about decision-making rely on the same mechanisms as having mental states and making decisions.

This first difficulty is aggravated by the second. It is not at all clear how to determine which brain regions are responsible for the "decision-making system– the same system that normally operates on natural non-pretend beliefs and desires" (Gallese and Goldman 1998). Until the defenders of off-line simulation provide a more concrete definition of "operates" and of "decision-making," I think the predictions of this model can neither be confirmed nor falsified by neuroimaging.

4 Conclusions

Theories of theories of mind have traditionally been tested with the tools of the philosopher, the developmental psychologist, the ethologist and the cognitive neuropsychologist. More recently, a new tool has become available in the form of functional magentic resonance imaging, but the potential theoretical contribution of fMRI remains relatively untested. In this chapter I have provided an illustration both of the possibility for theoretical impact (section 2) and the limits of this impact (section 3). The best way to distinguish between theoretical alternatives may still be through theoretical arguments (see Appendix).

Appendix: Theoretical arguments about off-line simulation

Philosophical arguments in favour of off-line simulation claim that simulation accounts for three features of inferences about other minds. The first intuition is that off-line simulation captures the immediacy of being drawn into someone else's subjective (emotional) experience. For instance, watching a rock-climber struggle before a fatal fall, Ravenscroft recounts: "Looking on, I vividly experienced what it was like to be him, and not only because, as a climber, I had been in similar situations before; any non-climber looking on could also experience what it was like to be that poor soul." (Ravenscroft 1998, p.171). It is the vividness, the emotionality, of his experience that Ravenscroft thinks is beyond the scope of any cold, detached (see Gordon 1998) knowledge- and inference- based account of theory of mind. Perhaps more compelling, empathy and perspective-taking (often taken to be fundamental components of theory of mind) can occur automatically, unintentionally or even against the will of the observer (Hodges & Wegner, 1997).

The second argument in favour of off-line simulation is that the causal principles used in reasoning about other minds are notoriously difficult to state precisely. "Why" challenges Goldman (1989) "should it be so difficult to articulate laws if we appeal to them all the time in our interpretative practice?"

Third and most important, off-line simulation provides an account of the actual mechanics of inferences about other minds, and especially of the integration of specific mental state contents into these inferences. Heal (1998a) defines content as the "representational aspect of a mental state, that in virtue of which it carries some specification of how the world is." Heal (1998b) takes the central question of theory of mind to be: "What is involved in our arriving at further psychological judgements about others given information about some of their existing psychological states? For example, provided with information about some of a person's beliefs and interests ('She believes p1 – p6 and is interested in whether q'), I may well arrive at a view about some likely further beliefs ('She believes that q')." To answer this question, Heal argues (1998b), a knowledge-based account of theory of mind is not merely empirically or currently insufficient; it is inapplicable a priori. We may believe a causal principle like 'beliefs and desires cause action', she concedes, but could we plausibly have a principle or generalisation about what actions or other beliefs the belief *that today is Tuesday* might cause? Stich and Nichols (1995) offer another example. Sven believes all Italians like pasta. Sven is introduced to Maria and told she is Italian. What will Sven say about Maria's feelings about pasta? Again, rather than reasoning about deduction-in-the-mind-of Sven, the intuition is that we arrive at Sven's conclusion by simulating his thinking in our own mind. Heal (1998a) concludes that "our primary competence with content is of the 'know how' variety and that only a small part of this can be reflected in any theoretical 'know that' about how contents relate." To make predictions about the causal role of specific beliefs we must make use of our own implicit procedural knowledge about inference – we must run an off-line simulation.

There is one further intuition in favour of off-line simulation, which is that people do seem to make competent inferences about the mental states of others in circumstances in which it seems implausible that they have any knowledge of the causal principle or generalisation at work. An example is provided by Kahneman and Tversky (described in Gallese and Goldman 1998). Two men in taxis are caught in a terrible traffic jam on the way to the airport, and each arrives half an hour after the scheduled departure time of his flight. The first man's flight left on time. But the second man discovers upon arrival that his flight was delayed, and is just this minute rolling away from the gate. How frustrated is each of these men? Our universal instinctive judgement is that the second man is more frustrated, having come, however accidentally, so close to making his flight. In this case my intuition is that we remember or simulate an emotional response to near misses and attribute a similar response to other people without ever representing *that* people tend to be especially frustrated by a near miss. It is only when brought to our attention by Kahneman and Tversky that this feature of the human experience becomes part of the content of our theory. In these cases, if indeed we do not have any other information

available (this is an empirical point, and not yet established, but at least intuitively possible), then we must use ourselves as an off-line model to predict how others will feel. Our ability to do so, and to do so effectively and accurately, would present a serious challenge to any purely knowledge-based account of inferences about mental states.

These thought experiments and philosophical considerations provide considerable initial plausibility to the off-line simulation account of behaviour explanation and predictions. However, I am not convinced.

A Emotional vividness

Ravenscroft's example, described above, of watching a tired man struggle for his life, illustrates an undeniable truth about social cognition: that we sometimes actually experience the emotions that we attribute to others. This phenomena of personal emotional engagement is clearly apparent while watching a film or reading a novel. Perner (1992) noted the power of imagined situations (e.g. that that man behind me is actually following me home) to provoke real emotions.

While this phenomenon is compelling and intriguing, it does not constitute a strong argument in favour of off-line simulation. The central feature of an off-line simulation is that the attributor's own "decision-making system" is run *off-line*, and the outputs are safely transferred to the "behaviour –predicting and –explaining system" instead of to the "action control system." Nobody incorrectly assigns someone else's *belief* to herself (e.g. coming to believe that *I am about to fall off this cliff*). So off-line simulation, as much as any other account, will have to provide an explanation of the special potency and contagion of emotions, that is distinct from the general architecture for attributing mental states. (There is also another, deeper problem concerning how people know which emotions to attribute to others in the first place, and I will address this problem below.)

B Articulating rules

The second claim is that the off-line simulation account provides the best explanation of the difficulty people (even experts, like the philosophical opponents of simulation) have when trying to articulate the causal principles or generalisations that guide theory of mind inferences. Off-line simulation explains this difficulty as a lack of knowledge: people cannot articulate these rules because we do not know them. (Recall that on the simulation account people have purely procedural knowledge about how minds work).

However, this explanation – that our knowledge of how minds work is purely procedural (or the related alternative: tacit) – seems to throw the baby out with the bath

water. Knowledge of mental causal relationships may be hard to articulate but it possesses all of the chief characteristics of ordinary, propositional knowledge (the following analysis is heavily indebted to Maibom 2003). Ordinary beliefs give rise to a characteristic conscious experience when the believer is confronted with an expression of the content of their belief (Stich 1978). Furthermore, these ordinary beliefs depend on concepts that have generality built into them (Davies 1989). Finally, ordinary beliefs enter into inferential relations with other beliefs. Thus if a person believes that the moon is made of cheese, the sentence "the moon is made of cheese" should give rise to this characteristic conscious experience. Still, we would not give him credit for a belief that the moon is made of cheese unless he could also conceive of other objects that he knows of being made of cheese. And last, if he subsequently learned that "all cheese is green," then his beliefs would be apt to give rise to the new belief that "the moon is green."

Knowledge of how minds work possesses all of these characteristics of ordinary knowledge. We easily and consciously recognise the formulations of mental causal principles that approximate our beliefs about how other minds work (e.g. people tend to act to achieve their desires given their beliefs). Psychological inferences are conceptual, in that they depend on genuine possession of the relevant psychological concepts, like 'belief' and 'desire.' (Compare language use, which proceeds beautifully even when the speaker does not possess the relevant linguistic concepts, such as 'dative' or '*wh*-movement'). And finally, we can acquire new psychological concepts (like "emotional intelligence," "hypnotised" and "suppressed desire"), which are integrated inferentially with the existing structure of our psychological knowledge, so that we can explain an apparently irrational behaviour as the product of certain beliefs (an old concept) and a suppressed desire (a new concept).

But if knowledge of the minds is ordinary knowledge, how *should* we account for the difficulty in articulating our common knowledge about the principles and generalisations that govern the mind? I am sympathetic to Maibom's (2003) analysis of the issue. Theories are not represented in the mind as a list of propositions, rules or generalisations; rather, theories (especially scientific theories) are more like models. Consequently, it is not surprising that people are not very good at translating their knowledge into these forms of representation. "But the point is that the generalisations badly express what subjects know, not that what subjects know is a bad theory," (Maibom 2003) or *a fortiori,* that subjects actually know nothing at all.

C Specific content

So neither the immediacy of theory of mind inferences, nor the difficulty that people have in enunciating the rules that govern those inferences, is a strong argument in

favour of off-line simulation. The third intuition is that simulation provides the best available account of how theory of mind inferences actually proceed – and in particular how the specific contents of mental states are integrated into those inferences. It is true that off-line simulation has a natural and ready-made account of the role of mental state contents: specific beliefs and desires are provided as inputs to the simulation, and the result is a specific action prediction. Alternative knowledge-based proposals, by contrast, have tended to focus on abstract causal relationships between kinds of mental states, the kind of structure Wellman and Gopnik (1992) call a 'framework theory'. Supporters of simulation are therefore right to criticise their opponents for lacking a sufficient explanation of how specific contents are integrated with these abstract kinds[17]. However, the account that off-line simulation does provide reveals some serious, and perhaps fatal, weaknesses in the off-line simulation proposal.

Some examples of the kinds of inferences about specific contents that a theory of mind should be able to explain were described above, including Sven's reasoning about Maria the Italian pasta lover (Stich and Nichols 1995). A similar problem is how we can answer the question "which city does Bill Clinton think is the capital of New York state?" (Nichols et al 1998). It is easy for me to attribute to Clinton the belief that Albany is the capital, and equally clear that I am not using a set of rules or causal principles concerning Clinton's beliefs about state capitals to do so. If I were using off-line simulation, I would imagine having Clinton's background beliefs and desires, and then imagine wanting to know which city is the capital of New York, and use the output provided by my own decision making system to predict Clinton's answer: Albany.

The problem with the off-line simulation procedure is made apparent, though, if I do not myself know which city is the capital of New York. In that case, I will not be able to predict Clinton's answer – and in fact my decision-making system will not be able to make any contribution at all towards a correct answer, since I cannot provide it with the relevant inputs for simulating Clinton. And yet, it seems obvious that the contribution of my theory of mind is not affected by my own knowledge or ignorance. The contribution

---

[17] Kelley's (1967) ANOVA model for the implementation of theory of mind inferences is an exception, in that he provides a worked example of how a specific inference – about why the man laughed at the comedian – could proceed. According to the ANOVA model, observers must collect a series of data: occasions on which the same man watched a different comedian, when different men watched the same comedian, and when the same man watched the same comedian on a different day. In this way the observer, like an experimental scientist, could determine which factor most reliably predicted the occurrence of the target behaviour. But manifestly, real life theory of mind inferences do not depend on such systematic data, which people do not, and often could not, collect. Instead theory of mind inferences, like many intuitive judgements, can be made given strikingly little data, usually on the basis of a single or a very few instances.

of my theory of mind, in both cases, is that Clinton thinks the capital of New York is the capital of New York[18], and I reach this conclusion based on facts I know about knowledge and about expertise, as well as other facts I know about Clinton, like his domain of expertise. My own knowledge plays a role only when I come to check which city *is* the capital of New York.

The same analysis can be applied to Sven[19]. My theory of mind contributes the prediction that Sven will make the rational *modus ponens* deduction about Maria, based on things I know about rationality and deductive inference and the consistency of beliefs. If the problem were an immensely complicated mathematical theorem, and Sven a professor of mathematics, my conclusion would be the same. My ability to make deductive inferences myself is useful only if I need to determine what the rational deduction would be; but the role of the theory of mind here is unaffected by whether I am either competent or inclined to derive the right conclusion.

So off-line simulation does not seem to provide the right story about how the specific contents of mental states are related to theory of mind inferences about those states. But that is not the worst of it. The recipe I provided above for simulating Clinton's knowledge about state capital reveals another flaw in the off-line simulation proposal. The first step in the procedure is to provide the right mental state inputs to my decision-making device so that I can accurately simulate Clinton. This may seem relatively undaunting, since very few of the differences between Clinton and I, or between our respective situations, are relevant for the inference at hand. Gordon (1998) sets the stage for a more difficult theory of mind inference. Hermia is in the forest. She has just awoken to find her lover Lysander missing, and his bitter rival Demetrius, asleep in his place. Gordon instructs Hermia that in order to simulate Demetrius' states of mind successfully, she should "transport herself in imagination into his situation to the extent to which it seemed, to a first approximation, relevantly different from her own [or] rather a self transformed, insofar as seemed necessary, into someone who would behave as she had known Demetrius to behave." But how?

---

[18] What my answer would be, if I was asked "What city does George W. Bush think is the capital of New York?," is a separate problem, raising the interesting question of the role of individual (and group) differences in theory of mind inferences. However, I will not address that question here.

[19] This argument also applies to a third, related thought experiment proposed by Harris (1992): how can one undergraduate predict the grammaticality judgements that another undergraduate will make about a set of sentences and pseudo-sentences in English, given that neither of them possess the relevant theory of linguistics that explains these judgements? Of course, the undergraduate need only predict that "other speakers of English will judge grammatical sentences to be grammatical" and then, if necessary, evaluate for himself which of the experimental stimuli fit the bill.

As I have mentioned previously, some versions of the off-line simulation architecture come with a special device – a dedicated pretend-state generator – responsible for generating accurate and relevant mental state inputs. And yet, looked at face on, this seems to be an extraordinary act of passing the buck. Because if the pretend-state generator can determine which are the accurate and relevant mental states of the target person, then why bother with simulation at all? Or rather, the actual off-line simulation seems to be limited to the role of hypothesis confirmation, checking whether the pretend-state generator's predictions about behaviour were accurate. The process of hypothesis checking may be interesting, and even critical, but is certainly not the all-encompassing architecture of action prediction and explanation that off-line simulation initially seemed to promise. Moreover, it seems likely that the only way the pretend-state generator could generate accurate and appropriate pretend states – since it obviously cannot use off-line simulation – would be if the generator incorporated some knowledge of how minds work, the very knowledge that off-line simulation aimed to do without

D The argument from error

In all, off-line simulation is in trouble. None of the three main intuitions in favour of this model turns out to make a strong case in favour of off-line simulation, and at least in the case of specific content, the intuition goes clearly in the opposite direction. There is one more argument against the sufficiency of off-line simulation to account for theory of mind inferences that I find compelling: the argument from error.

When psychological prediction goes smoothly it may be hard to distinguish between off-line simulation and knowledge-based accounts of the process. But when observers predict incorrectly, Stich and Nichols (1995) argued, off-line simulation has only two explanations available. Either (i) the observer and actor have different practical reasoning systems, or (ii) the inputs given to the observer's system were not accurate copies of the actor's beliefs and desires. Consequently, if observers make systematic mistakes when neither of these options are plausible, off-line simulation will be in trouble; especially if those mistakes correspond to biases in the observer's explicit knowledge about how minds work. This is the argument from error.

The clearest example of a systematic difference between how adults behave and how they think they behave is in the domain of rational decision making. In both statistical inferences and risky choices, people's intuitive choices depart from normative theories of probability and utility (Shafir and Tversky 1995). Critically, people's predictions of how they will behave, and expectations of how they should behave, correspond more closely to the normative theory, than to their subsequent behaviour. "The axioms of rational theory are so intuitively appealing that many have considered it

likely that a theory derived from these axioms would provide an acceptable, if somewhat idealised, account of actual behaviour," writes Shafir (1993). "Errors of reasoning are often considered embarrassing because there is a feeling that we could, and should, have done better." Indeed, exposing human irrationality is a cottage industry in psychology, and the charisma of this research program derives precisely from the fact that such irrationality violates our expectations and challenges our theory of ourselves.

The defender of off-line simulation may reply to this challenge, that when people are asked to make predictions about rational choice, for instance, the simulated inputs do not accurately reflect the inputs of a person asked to make the choice. To counter this response Nichols et al (1998) actually performed a relevant experiment[20]: half of the

---

[20] Perner et al (1999) also claimed to have conducted a relevant – indeed, the critical – experiment to test this prediction of off-line simulation. However, I find both the motivation and the results of this experiment to be muddled and hard to interpret.

Perner et al (1999) asked participant subjects to make simple magnitude judgements ("how fast is a good race horse?") either on a short scale (0 to 90 km/h) or on a long scale (0 to 130 km/h). Predictably, subjects who were given the long scale tended to judge that a good racehorse is faster than the subjects given the short scale did. When the observer subjects were given both scales simultaneously, in every case the predicted discrepancy was as big as, or bigger than, the actual frame effect on participants, and was bigger than the frame effect predicted by observer subjects given only one scale.

My interpretation of these results – though not Perner's – is that the juxtaposition of the two scales highlighted the experimental variable, the frame. Even subjects who did not begin the experiment with knowledge of frame effects could have inferred the intention of the experimenter pragmatically, and so the causal principle in human cognition under investigation. Subjects therefore predicted that participants' judgements would follow the general causal principle of frame effects for all of the stimulus items – even those where the participants themselves escaped the bias. This interpretation also makes sense of the results of Perner et al's Experiment 3: juxtaposition of a scale and a free-rating question ("How fast is a good race horse? ___ km/h") did not lead to systematic predictions. In fact juxtaposition had very little effect in this case, leading to judgements very similar to those made by observer subjects given either one response format or the other. The difference between a scale and a free-response is not highly salient, as it is with the two scales, and so subject may have been unable to infer the intentions of the experimenter, or the cognitive principle in question. My interpretation could be tested by asking observer subjects in each condition how they made their predictions, but Perner et al did not do this.

The most confusing feature of Perner et al's experiment is that observer subjects given a single scale sometimes differed significantly in their predictions from the performance of actual participants. I cannot think of a model of theory of mind inferences that would predict different responses to the questions "Please estimate the speed of a good racehorse and mark your estimate ..." and "If a person was asked to estimate the speed of a good racehorse, what would this person mark on the scale below?" Given no information about the cognitive properties of "this person," the best possible response is to reason that the person will be as accurate as possible, and so their mark will be close to whatever the speed of a racehorse is, and the best approximation of that response that I could possibly achieve would be to mark whatever *I* think that speed would be. Nevertheless, some of Perner et al's observer subjects, presented with only one scale, appeared to escape the frame effect. I can think of no explanation of this result.

subjects were asked to make a simple judgement, and the other half of the subjects – in as similar a procedure as possible – were asked to predict what judgement another person would make in that circumstance. The observer subjects' predictions deviated systematically from the actual judgements of the participant subjects.

Nichols et al's (1998) demonstration relied on the Langer (1975) effect: subjects who have chosen a lottery ticket for a random draw are willing to sell it only for a significantly higher price (six to eight times higher) than if they had been merely given the ticket. The asking price predicted by observer subjects, who watched a participant go through the procedure on a video, did not depend on the contrast between asking and giving, but instead remained closer to the normative value of the ticket in terms of the probability and value of winning the lottery. Since subjects were drawn from the same undergraduate pool, and divided randomly into the participant and observer groups, it seems highly unlikely that these two groups have systematically different practical reasoning systems. Moreover, observer subjects were given every possible opportunity to simulate the inputs of the participant subjects. If an observer ran an off-line simulation on these inputs, then the irrationality instantiated in his human decision-making system would be manifest in his prediction. The fact that the observers' predictions followed the normative theory and over-rationalised the expected behaviour, rather than the actual behaviour, suggests that subjects were using some knowledge or theory of human behaviour – including an over-application of the principle of rationality.

The standard off-line simulation rebuttal to this kind of evidence is to say that still, even with the video presentation, observers may not encode the inputs of the situation precisely enough to perform an accurate simulation. Goldman (1992), for instance, suggests that "the simulated scene is unlikely to be as detailed as the a real perceptual scene; one is unlikely to replicate the uncertainties of the live situation; and one is unlikely to simulate in detail the actual scanning and comparison operations. Thus there is no reason to expect that a simulation would generate the same behaviour as the actual incident." If this is the case with a full video presentation, how could a simulator in the highly unspecified conditions of a normal theory of mind inference ever hope for accuracy?

The argument from error is even more powerful when behaviour predictions depend on an immature theory of mind, which includes more striking departures from the way minds actually work. Ruffman (1994, described by Perner 1998) provides very elegant evidence that children are reasoning with a (faulty) theory about other people's knowledge. A child and an adult observer ('A') are seated in front of two dishes of beads. The round dish contains red and green beads, while the square dish contains only yellow beads. Both A and the child watch while a bead from the round dish is moved under

cover into an opaque bag. The child, but not A, knows that the chosen bead was green. Then the child is asked "what colour does A think the bead in the bag is?" Overwhelmingly, the children report that A thinks the bead is *red.*

How can we explain this? If the child was simulating A, we would expect either that the child would accurately express A's ignorance, or else would assimilate A to the self and judge that A thinks the bead is green, but this did not happen. Nor we can we claim the child simply misrepresented or mis-remembered what A saw, and so ran a faulty simulation: no one reports that A thinks the bead is yellow. The actual result is best explained by an inaccurate generalisation in the child's developing theory of mind: "ignorance means you get it wrong." Since A is ignorant of which bead was chosen from the round dish, A must think that it was the wrong colour, a red one.

The two experiments described above illustrate the argument from error: both children and adult observers make systematic errors in their predictions of others' beliefs and behaviour, exactly when their theory of human behaviour includes inaccuracies. These cases cannot be explained by off-line simulation. Greenwood (1999) has therefore suggested that off-line simulation should give up trying to account for conscious action explanation and prediction, and focus instead on (tacit) social *anticipation.* If off-line simulation is designed merely to anticipate the future actions of others, then these explicit tests of verbal behaviour prediction may underestimate the power and accuracy of simulation, by testing the wrong response modality. Even if this manoeuvre worked, though, it would be a Pyrrhic victory. Explicit action explanation and prediction lie at the core of our human theory of mind.

References

Brass, M., H. Bekkering, et al. (2001). Movement observation affects movement execution in a simple response task. *Acta Psychol (Amst)* 106(1-2): 3-22.

Buccino, G., F. Binkofski, et al. (2001). Action observation activates premotor and parietal areas in a somatotopic manner: an fMRI study. *Eur J Neurosci* 13(2): 400-4.

Buxbaum, L. J. (2001). Ideomotor apraxia: a call to action. *Neurocase* 7(6): 445-58.

Carey S (1985) *Conceptual change in childhood.* MIT Press

Carey S and Spelke E (1994) Domain-specific knowledge and conceptual change. In *Mapping the Mind: Domain specificity in cognition and culture.* L. Hirschfeld and S. Gelman. (eds) New York, NY: Cambridge University Press.

Chaminade, T., A. N. Meltzoff, et al. (2002). Does the end justify the means? A PET exploration of the mechanisms involved in human imitation. *Neuroimage* 15(2): 318-28.

Chomsky N (1980) Rules and representations. *The behavioral and brain sciences.* 3:1-61

Decety, J., J. Grezes, et al. (1997). Brain activity during observation of actions. Influence of action content and subject's strategy. *Brain* 120(Pt 10): 1763-77.

Decety, J., M. Jeannerod, et al. (1989). The timing of mentally represented actions. *Behav Brain Res* 34(1-2): 35-42.

di Pellegrino, G., L. Fadiga, et al. (1992). Understanding motor events: a neurophysiological study. *Exp Brain Res* 91(1): 176-80.

Flavell (1988) The development of children's knowledge about the mind: From cognitive connections to mental representations. In *Developing theories of mind* JW Astington, PL Harris et-al (eds) New York, NY: Cambridge University Press.

Fletcher, P. C., F. Happe, et al. (1995). "Other minds in the brain: a functional imaging study of "theory of mind" in story comprehension." *Cognition* **57**(2): 109-28.

Fodor (1983) *The modularity of mind.* Cambridge MA: MIT Press.

Fogassi, L., V. Gallese, et al. (1996). Coding of peripersonal space in inferior premotor cortex (area F4). *J Neurophysiol* **76**(1): 141-57.

Fogassi, L., V. Raos, et al. (1999). Visual responses in the dorsal premotor area F2 of the macaque monkey. *Exp Brain Res* **128**(1-2): 194-9.

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

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

Gallese, V., L. Fadiga, et al. (1996). Action recognition in the premotor cortex. *Brain* **119**(Pt 2): 593-609.

Gangitano, M., F. M. Mottaghy, et al. (2001). Phase-specific modulation of cortical motor output during movement observation. *Neuroreport* **12**(7): 1489-92.

Gibson (1966) *The senses considered as perceptual systems*. Oxford, England: Houghton Mifflin.

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

Gopnik, A. (1996) The scientist as child. Philosophy of Science. 63:485-514

Gopnik, A. and Wellman HM (1994) The theory theory. In *Mapping the Mind: Domain specificity in cognition and culture*. L. Hirschfeld and S. Gelman. (eds) New York, NY: Cambridge University Press.

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

Gopnik, A. and J. W. Astington (1988). Children's understanding of representational change and its relation to the understanding of false belief and the appearance-reality distinction. *Child Dev* 59(1): 26-37.

Gordon, R (1998) Radical Simulation. In *Theories of Theories of Mind*, Ed. Carruthers & Smith

Greenwood J.D. (1999) Simulation, Theory-Theory and Cognitive Penetration: No 'Instance of the Fingerpost' *Mind & Language*, 14(1) pp. 32-56

Grezes, J. and J. Decety (2001). Functional anatomy of execution, mental simulation, observation, and verb generation of actions: a meta-analysis. *Hum Brain Mapp* **12**(1): 1-19.

Halsband, U., J. Schmitt, et al. (2001). Recognition and imitation of pantomimed motor acts after unilateral parietal and premotor lesions: a perspective on apraxia. *Neuropsychologia* 39(2): 200-16.

Hanna-Pladdy, B., K. M. Heilman, et al. (2001). Cortical and subcortical contributions to ideomotor apraxia: analysis of task demands and error types. *Brain* 124(Pt 12): 2513-27.

Harris P (1989) *Children and emotion: The development of psychological understanding*. Cambridge, MA: Basil Blackwell, Inc

Harris P (1992) From simulation to folk psychology: the case for development. *Mind and Language* 7, p. 120- 144

Heal J (1998) Simulation, theory, and content. In *Theories of Theories of Mind*, Ed. Carruthers & Smith

Heal J. (1998) Co-Cognition and Off-Line Simulation: Two Ways of Understanding the Simulation Approach *Mind & Language* 13(4): 477-498

Hodges S & Wegner D (1997). Automatic and Controlled Empathy In W.J. Ickes (Ed). *Empathic accuracy.*

Iacoboni, M., L. M. Koski, et al. (2001). Reafferent copies of imitated actions in the right superior temporal cortex. *Proc Natl Acad Sci U S A* **98**(24): 13995-9.

Kareken, D. A., F. Unverzagt, et al. (1998). Functional brain imaging in apraxia. *Arch Neurol* 55(1): 107-13.

Kelley HJ (1967) Attribution theory in social psychology. In D Leving (Ed) *Nebraska Symposium on Motivation* Vol 15.

Kohler, E., C. Keysers, et al. (2002). Hearing sounds, understanding actions: action representation in mirror neurons. *Science* **297**(5582): 846-8.

Langer (1975) The illusion of control. *JPSP* 32(2): 311-28.

Leslie, A. (1994). A theory of ToMM, ToBy, and Agency: Core architecture and domain specificity. *Mapping the Mind: Domain specificity in cognition and culture*. L. Hirschfeld and S. Gelman. New York, Cambridge University Press: 119-148.

Leslie, A. (2000). 'Theory of Mind' as a mechanism of selective attention. *The New Cognitive Neurosciences*. M. Gazzaniga. Cambridge, MA, MIT Press: 1235-1247.

Maibom H. (2003) The Mindreader and the Scientist *Mind & Language*, 18(3) pp. 296-315

Malle B (1999) How people explain behaviour: A new theoretical framework. *Personality and Social Psychology Review* 3(1) 23-48.

Marr (1982) *Vision*. Cambridge, MA: MIT Press.

Murata, A., V. Gallese, et al. (1996). Parietal neurons related to memory-guided hand manipulation. *J Neurophysiol* 75(5): 2180-6.

Nichols S, Stich S, Leslie A & Klein D (1998) "Varieties of offline simulation" in *Theories of Theories of Mind*, Ed. Carruthers & Smith

Perner J (1991) *Understanding the representational mind.* Cambridge, MA: The MIT Press.

Perner J (1998) "Simulation as explication of predication-implicit knowledge about the mind: arguments for a simulation-theory mix" in *Theories of Theories of Mind*, Ed. Carruthers & Smith

Perner J and Howes D (1992) "He thinks he knows": And more developmental evidence against the simulation (role taking) theory. *Mind-and-Language.* 7(1-2): 72-86

Perner J. Gschaider A. Kühberger A. Schrofner S. (1999) Predicting Others Through Simulation or by Theory? A Method to Decide *Mind & Language* 14(1) 57-79

Povinelli DJ and Giambrone S (2001) Reasoning about beliefs: a human specialization? *Child Dev.* 72(3):691-5.

Ravenscroft I (1998) What is it like to be someone else? Simulation and empathy. *Ratio* XI p 170 – 185.

Rizzolatti, G., L. Fadiga, et al. (1996). "Premotor cortex and the recognition of motor actions." *Brain Res Cogn Brain Res* 3(2): 131-41.

Rizzolatti, G., L. Fadiga, et al. (1999). "Resonance behaviors and mirror neurons." *Arch Ital Biol* **137**(2-3): 85-100.

Rizzolatti, G., L. Fogassi, et al. (2001). "Neurophysiological mechanisms underlying the understanding and imitation of action." *Nat Rev Neurosci* **2**(9): 661-70.

Rizzolatti, G., M. Gentilucci, et al. (1990). "Neurons related to reaching-grasping arm movements in the rostral part of area 6 (area 6a beta)." *Exp Brain Res* 82(2): 337-50.

Roth D and Leslie A (1998) Solving belief problems: Toward a task analysis. *Cognition* 66(1): 1-31

Scholl B and Leslie A (1999) Modularity, development and 'theory of mind.' *Mind-and-Language.* 14(1): 131-153

Shafir E (1993) Intuitions about rationality and cognition. In *Rationality: Psychological and Philosophical Perspectives.* KI Manktelow and DE Over (eds). London UK: Routledge.

Shafir E and Tversky A (1995) Decision making. *In An Invitation to Cognitive Science vol. 3 Thinking.* Smith E and Osherson D (eds). Cambridge MA: MIT Press

Stich S & Nichols S (1995) "Second thoughts on simulation" In Stone & Davies (Eds) *Mental simulation: evaluations and applications.* Oxford.

Stich A & Nichols A (1998) Theory theory to the Max. *Mind and Language* 13(3):421-449.

Stich S (1978) Beliefs and subdoxastic states. *Philosophy of science* 45, p 499 – 518.

Strafella, A. P. and T. Paus (2000). "Modulation of cortical excitability during action observation: a transcranial magnetic stimulation study." *Neuroreport* 11(10): 2289-92.

Sturmer, B., G. Aschersleben, et al. (2000). "Correspondence effects with manual gestures and postures: a study of imitation." *J Exp Psychol Hum Percept Perform* 26(6): 1746-59.

Toraldo, A., C. Reverberi, et al. (2001). "Critical dimensions affecting imitation performance of patients with ideomotor apraxia." *Cortex* 37(5): 737-40.

Umilta, M. A., E. Kohler, et al. (2001). "I know what you are doing. a neurophysiological study." *Neuron* **31**(1): 155-65.

Vogeley, K., P. Bussfeld, et al. (2001). "Mind reading: neural mechanisms of theory of mind and self-perspective." *Neuroimage* **14**(1 Pt 1): 170-81.

Wohlschlager, A. and H. Bekkering (2002). "Is human imitation based on a mirror-neurone system? Some behavioural evidence." *Exp Brain Res* **143**(3): 335-41

Wolf, N. S., M. Gales, et al. (2000). Mirror neurons, procedural learning, and the positive new experience: a developmental systems self psychology approach. *J Am Acad Psychoanal* 28(3): 409-30.