

Theory of Mind and the Evolution of Social Intelligence

Valerie E. Stone

School of Psychology

University of Queensland

In press: in *Social Neuroscience: People Thinking About People*, J. Cacciopo, Ed., MIT Press

29 pages w/o refs and figures

Mailing address:

School of Psychology, University of Queensland

McElwain Building

St. Lucia, Queensland 4072

AUSTRALIA

Phone: +61 7 3346 9517

Fax: +61 7 3365 4466

Email: v.stone@psy.uq.edu.au

Web page: <http://www.psy.uq.edu.au/~stone>

Social Intelligence and Domain-Specificity in an Evolutionary Perspective

I would like to place social cognition and thus the social brain, in an evolutionary context. Humans are social animals, adapted to living in groups. Group living probably goes back at least 54 million years in our family tree, to our common ancestor with other primates (Yoder et al., 1996; Foley, 1997). Many of our social behaviors are shared with our primate cousins, and so it is likely that many of our social cognitive abilities are as well. On the other hand, each species has its uniqueness. Humans, unlike other primates, use language, plan for the future and make complex inferences, and all of these abilities affect our social behavior and cognition. These abilities, however, may depend on a relatively small number of uniquely human capacities, with most of our social cognition being in common with our primate cousins. Recent studies in evolutionary biology and primatology have revealed surprising cognitive capacities in great apes -- orangutans, with whom we share a common ancestor 14 million years ago (mya), and chimpanzees and bonobos, with whom we share a common ancestor 5-7 mya (Foley, 1997, Gibbons, 2002; see Figure 1). African and Asian monkeys (common ancestor 26 mya) for the most part do not seem to share these abilities, but apes have been found to engage in tool use, cultural learning, and insightful problem-solving (Suddendorf, 1999; McGrew, 2001). These discoveries have forced psychologists and neuroscientists to define more narrowly what is unique about human minds, and have given us a basis for understanding how our complex cognition can be continuous with that of our closest relatives and continuous with our hominid ancestors.

Social behavior can be defined as any interaction with members of one's own species. Social cognition, then, is the information-processing architecture that enables us to engage in social behavior. Social neuroscience is the study of how the brain implements the information-processing architecture for sociality. A key question for social neuroscience is to what extent brain systems subserving social behavior are *socially specific*, consisting of neural processes that

operate only on social information, and to what extent these systems are more general, consisting of neural processes that subserve multiple areas of cognition. Research with neurological patients can give us insight into these questions, as we find evidence from dissociations either that social cognition can be impaired independently of more general cognitive processes, or that deficits in social cognition are always accompanied by more general cognitive impairments. Neuroimaging can reveal whether the brain systems involved in different kinds of social tasks are in the same or different locations in the brain, and how much those systems overlap with areas involved in more general cognitive functions.

I will argue that there are key aspects of our social cognition that depend not on specifically social processes, but on some very general and powerful cognitive abilities that are unique to our species, and that are used in many other contexts besides social cognition. Our capacity for executive control over cognition, for metarepresentation, and for recursion enable not only our complex social cognition, but many of our other uniquely human abilities as well¹: symbolic language, syntax, future planning, episodic memory, just to name a few (Suddendorf, 1999; Corballis, 2003). The remarkable changes that have taken place in hominid brain evolution have been due to expansion of our general cognitive capacities, particularly those subserved by the frontal lobes. One could argue that these general cognitive capacities were selected for in primate and hominid evolution *because* they were so useful for social behavior, indeed Brothers has made such an argument for the general perceptual computations that subserve face recognition (Brothers, 1997). Such arguments are historical speculations rather than scientific theories – it is difficult to know what selection pressures operated in the past. One can make some inferences from the adaptive design of a cognitive system, arguing that a cognitive system will perform the function it was selected for most efficiently, and that it will have design features specific to that function (Tooby & Cosmides, 1992). However, executive function and recursion

clearly fail that test, as they are equally useful in memory, language, social cognition and tool manufacture, and have no design features specific to the social domain (Corballis, 2003).

Much of our social behavior, however, *is* similar to that of other primates and other mammals. It is parsimonious to assume that many of the brain systems that subserve social behavior in humans and in other social mammals with whom we share common descent are inherited from a common social mammal ancestor, and will share common features. The use of animal models in social neuroscience depends on such assumptions. In comparative biology, this is known as *homology* to indicate that two species share similar structures because of common descent and common genetics, not because it evolved independently in two different lines. If forming a mother-infant bond looks similar in many branches of our family tree, then we can reasonably expect that the brain systems that implement that behavior will be similar in those branches. Furthermore, even our higher level social cognitive abilities, those that depend on language and a complex self-representation, should be seen as continuous with those of our primate ancestors, in whom the building blocks of our complex sociality and cognition first emerged.

Although some domain-specific brain systems for sociality may have been preserved for tens of millions of years, social neuroscientists should be wary of assuming too much social specificity in social cognition. With each domain of social cognition – attachment, hierarchy negotiation, cooperation, in-group/out-group categorization – one needs to define carefully what the components of an ability are, consider whether each component is an aspect of cognition that we share with other social primates, and then do careful work to define whether each component is specific to social behavior or is used in non-social contexts, and whether it depends on more general cognitive processes.

Social cognition in monkeys, apes and humans: Components of Theory of Mind

As an example of how to approach a topic in social neuroscience in this way, I would like to review research on our ability to understand other people's mental states, a cognitive capacity known as "theory of mind" (ToM). Humans make inferences about and interpret others' behavior in terms of their mental states, meaning their emotions, desires, goals, intentions, attention, knowledge and belief. ToM thus encompasses a variety of cognitive processes, and takes several years to unfold in human development (Baron-Cohen, 1995, Wellman & Liu, 2004). By breaking ToM down into components, we can ask which of those components are shared with our ape relatives, which are uniquely human, and which seem to be socially specific.

First, however, we need to clarify some issues of terminology, as different scholars and different fields sometimes use the term "theory of mind" differently. The developmental and primate literature on theory of mind makes distinctions between several different types of mental state inferences. Before the age of 4, children can make inferences about others' intentions, goals, desires, wants and feelings. Somewhere between 3 and 4, children can infer what are called *epistemic* mental states: knowledge, belief, perception. ToM is sometimes used to refer specifically to the ability to represent the *contents* of one's own and others' mental states, something that younger children cannot do. Theory of mind is often seen as equivalent to *metarepresentation*, the ability to represent representations, as in "He thinks that [his car is in the garage]" or "She saw that [the lion had escaped from its cage]" (Leslie, 1987; Perner, 1991; Baron-Cohen, 1995; Suddendorf, 1999). Not all of the mental states that we routinely infer require metarepresentational inferences. For example, inferring another person's emotional state does not require representing someone else's representations, but only their external appearance, "She looks angry." For this reason, Leslie and Frith have argued that inferring emotional states should *not* be considered theory of mind (1990). Inferring others' intentions, goals and desires is another grey area for

theory of mind terminology, as such inferences also do not necessarily require metarepresentation. Children can make these inferences much earlier than they can do metarepresentation (Suddendorf & Whiten, 2001). In social neuroscience as in developmental psychology “theory of mind” is used broadly to mean inferring a variety of mental states, not limited to metarepresentation. It is important to be clear which type of mental state is meant when using the term theory of mind.

Developmental psychology can reveal the building blocks and components of theory of mind (ToM), and comparative ethology can tell us whether or not our mammal and primate relatives have each component (Saxe, Carey & Kanwisher, 2004; Suddendorf, 1999). The groundwork is then laid for neuroscience to investigate the brain systems involved in each component. Many building blocks of theory of mind are present in our ape relatives (Suddendorf, 1999; Suddendorf & Whiten, 2001; Hare, Call & Tomasello, 2001; Leavens, Hopkins & Thomas, 2004). Thus, those building blocks at least are more ancient than the hominid line of the past 5-7 million years.

[Insert Figure 1 about here]

Below, I discuss developmental studies on how humans develop ToM and relevant primate studies, before reviewing the neuroscience data on the brain systems involved. In reviewing the developmental literature, I remain agnostic on theories of *how* ToM develops, whether through simulation, theory-building, or modular maturation. Rather, I focus on when certain abilities emerge, and which abilities might share common processes. From such a review, it is clear that we have gaps in research on the neural basis of ToM. For example, although there is ample research on systems underlying our ability to detect eye gaze direction, there is almost no research on what we do with eye gaze, *i.e.*, following gaze or establishing joint attention. Much of the research on “theory of mind in the brain” does not include control conditions with equivalent

task demands (and I include some of my own research in this), and thus leaves open the question of whether there are areas involved *specifically* in ToM. As a corrective to this, I suggest taking a careful look at the components of ToM to see how future research might take account of these issues.

Building Blocks of ToM: Inferring Goals and Intentions

Infants from very early on begin to distinguish actions that are intentional, and to discern an actor's goal. Infants between 5 and 9 months can differentiate accidental from intentional behavior (Woodward, 1999), and by 15 months, infants classify actions according to the goal of an action (Csibra et al., 2003). These results show an implicit understanding of intentions and goals. Call and Tomasello (1998) showed that chimpanzees (*Pan troglodytes*) and orangutans (*Pongo pygmaeus*) could also distinguish visually between accidental and intentional actions. Monkeys do not seem to make this distinction. Assuming homology, this would put the date for this ability in a common ancestor at about 14 million years ago.

Jellema et al. (2000) have begun to investigate the neural networks involved in detection of goals. They recorded cells in the anterior STS of macaque monkeys that responded to the sight of an agent reaching for something, but only when the agent was looking at the point reached for, with "looking at" indicated by eye gaze, head orientation, and upper body orientation. They propose that these cells integrate input from cells that respond to gaze direction with cells that respond to limb movement direction (Jellema et al., 2000). Such an integration is a necessary part of the cognitive architecture of detecting goals, which is the first step in understanding intentional action. After apes branched off from monkeys, further elaborations of this architecture may have occurred, involving areas beyond STS.

Building Blocks of ToM: Joint attention

Between the first and second year, children treat others' gaze direction as a source of information, indicative of that person's focus of attention. In "joint attention," emerging between 18-24 months, the child takes an active step beyond gaze monitoring, to call adults' attention to particular objects, by pointing or holding up something for them to see. Establishing empirically that a child is using joint attention usually depends on clear evidence that the child has either moved an object deliberately into another person's line of view, or that they are using "protodeclarative pointing," i.e., pointing to something and alternating gaze between another person and the object (Baron-Cohen, 1995; Franco & Butterworth, 1996).

It is worth noting that children at this stage do not always *successfully* bring things to adults' attention. Children at this age make systematic errors about what others can and cannot see (Mossler, Marvin & Greenberg, 1976; Liben, 1978; Flavell et al., 1981), and call adults' attention to things the adult cannot see (such as holding up a watch and saying to a parent on the phone, "Mommy, look at my new Mickey Mouse watch!"). Thus, there is still no evidence that children at this stage can correctly understand the *contents* of others' mental states. The child's representation may only integrate information about whether or not an adult is paying attention (binary yes/no), the rough location of objects in space, and a repertoire of actions that generally succeed in engaging adult's attention (holding things up, pointing). Nevertheless, children at this age become more active in trying to affect others' attention.

What evidence is there for joint attention in primates? Kumashiro et al. (2002) have presented suggestive evidence that a Japanese macaque monkey (*macaca fuscata*) in their lab engaged in gaze monitoring and learned to use protodeclarative pointing. In a later study, the same research group found that monkeys that had been trained to learn this kind of joint attention could copy complex motions by human experimenters, whereas monkeys that did not

engage in joint attention did not copy a human experimenter (Kumashiro et al., 2003).

Chimpanzees and orangutans have clearly been observed to engage in gaze monitoring as a kind of visual joint attention. They respond differently to situations where another animal or human experimenter can see an object clearly, and situations in which the other's gaze is occluded (Hare, Call, Agnetta & Tomasello, 2000). Chimpanzees seem to use gestures differently when a human experimenter is or is not looking at a food item they desire (Povinelli, Theall, Reaux & Dunphy-Lelii, 2003). Apes, like monkeys, do not seem to use referential pointing spontaneously in their natural environment. Even apes in captivity usually use pointing to request an object (Povinelli & O'Neill, 2000). Given that in their natural environment, monkeys walk on all fours and apes knuckle-walk, and that they all rely on their hands for climbing, it is perhaps not surprising that they don't use their hands for such gestures, and instead rely on a cue that is easily available. Since the empirical standard for joint attention is pointing with a finger, there may be a somewhat human-centric methodological bias against finding evidence of joint attention in primates.

No patient or neuroimaging research to date has focused specifically on joint attention. Joint attention is an important stage in ToM development, thus this is a significant omission.

Building Blocks of ToM: Pretend play

At the same age, children begin to engage in pretend play (18-24 months). Pretending involves decoupling the pretend reality ("this is my baby") from perceptual reality ("this is an inanimate doll"). There is considerable debate in developmental psychology over what children understand about pretense as a mental state. Leslie (1987) argues strongly that pretense involves representing one's own and others' mental states, that is, that children have a representation such as "Mommy is pretending that → [the doll is a baby]." However, children at this age still fail perspective-taking tasks and make systematic errors about what others can and cannot see

(Mossler, Marvin & Greenberg, 1976; Liben, 1978; Flavell et al., 1981). Thus, it is difficult to make a convincing case that they can represent the contents of a playmate's mental states. Furthermore, younger children do not always understand the role of mental states in pretense, (Lillard et al., 2000). In debates over whether children at this age truly understand the representational nature of pretend play, the more parsimonious alternative hypothesis is that they treat pretense as a special kind of action (e.g., Wellman & Lagattuta, 2000). Neuroscience research into the brain substrates for pretense might help resolve this debate. If an understanding of pretense is just a special kind of action, then areas involved in pretense might overlap with areas involved in representing actions not currently being done, i.e., the supplementary motor area. If pretense does involve representing the contents of others' mental states, then the same areas active for ToM should be active for pretense. No neuroscience studies of pretense have yet been reported in the literature, so the question remains open.

ToM and implicit mentalistic understanding: Acting based on others' mental states

Something genuinely new emerges between ages 2 and 3. Children begin to demonstrate an understanding of some of the properties of mental things as opposed to physical things. They seem to understand that mental states such as desire and knowledge are private, unobservable directly, and can change or not change independent of reality. I will refer to this understanding as *mentalism*, to denote *understanding of the properties of mental things*. Wellman discusses this as "belief-desire psychology" (Wellman, 1990). Children at about age 3 also begin to demonstrate implicit knowledge of the *contents* of others' mental states, though not explicit knowledge. This aspect of ToM, mentalism, may be what is socially specific.

Desire. Beginning at around age 2, children readily use language about desire, e.g. "she likes," and seem to understand that people's attitudes and emotions towards various objects can be used to predict what they will do (Wellman & Lagattuta, 2000; Wellman *et al.*, 2001). Thus,

understanding desire may bootstrap off of an understanding of intentions/goals in development. At this age, children can also understand that different people's desires are distinct, that, for example, they don't want to eat their vegetables, but grownups seem to like this yucky-tasting stuff. Experimental evidence indicates that the ability to understand *diverse desires* emerges between 18-24 months (Wellman & Woolley, 1990; Repacholi & Gopnik, 1997; Wellman & Liu, 2004). Likes, wants and desires are private mental states that are changeable. A person's likes and wants can change independent of external reality changing. This developmental step represents the first stage at which children display *mentalism*, an understanding of something that is uniquely mental, private and decoupled from the external world.

Belief and knowledge. Adult human theory of mind involves understanding epistemic mental states, knowledge and beliefs. Children and primates may or may not fully understand these mental states. Knowledge is cumulative compared to desire, but like desire, knowledge is also changeable, and decoupled from the external world. Children do show an *implicit* understanding of the changeable nature of knowledge and belief before they can talk about it or understand it explicitly.

To test whether children or primates know about someone else's knowledge state, one has to distinguish their representation of someone else's mental state from their representation of the state of reality. If one probes what a subject thinks someone else knows, and what that person knows is true, it is always possible that the subject is just responding with what he/she knows him/herself. Thus, testing whether a subject can understand that someone else holds a false belief has long been held to be the key test of theory of mind. But what does it mean to understand false belief? Does someone understand it if he/she can act based on someone else's false belief, but can't talk about it, or does someone have to be able to talk or answer verbal

questions about it explicitly? Understanding false belief, it appears, can be either implicit or explicit.

Two basic kinds of false belief tasks have been used with children, *location change* tasks and *unexpected contents* tasks¹. In a location change task, the subject is told a short story (and shown pictures to go with the story, or the story is acted out with toy figures) in which character A puts an object in location 1, and then turns away or goes out of the room. Character B moves the object to location 2 while character A cannot see, and then the subject is asked where character A will look for the object, location 1 or location 2. Children generally can pass this test some time late in the 3rd year of life and age 4, but it is rare that 3-year olds can pass it (Wellman *et al.*, 2001).

However, 3-year olds can pass an implicit version of the task. Perner and Garnham (2001) used an ingenious test to demonstrate this. The child was told that another person was going to slide down one of two slides, and that the child was supposed to place a mat so the person could land safely at the bottom of the slide he or she was going to come down. The situation was manipulated so that sometimes the person who was supposed to slide down had a true belief about which slide they were supposed to come down, and sometimes had a false belief, and would therefore come down the wrong slide. The task was set up so that the children had to act quickly, without time for deliberation. The 36-month-olds in this study were likely to place the mat under the correct slide, showing that they had an implicit ability to track the other person's belief state, true or false. These same children failed a standard false belief task that asked for an explicit choice of which slide the person would come down.

Pratt & Bryant (1990) found that 40-month-olds could pass a "seeing-leads-to-knowing" test at an age when children generally cannot pass false belief tasks. They showed children two pictures, one of a girl looking into a box and one of a girl touching the box but looking away, and asked which one knew what was in the box. This task does not require explicitly reporting the

contents of the girl's mental state, and thus is a more implicit task than false belief tasks. Thus, an implicit understanding of the fact that knowledge can change independent of reality, and that such changes are linked to perception, seems to emerge around age 3.

Chimpanzees (*Pan troglodytes*) can do a task which may also reflect an implicit understanding of knowledge and ignorance. In chimpanzee society, if two animals see the same piece of food, the dominant male will almost always get it, and trouble follows if he does not. Chimpanzees in this task were given a choice to head towards one of two food items, one which a dominant animal could see, and one which he couldn't see, or one location in which he had seen food being hidden, but had not seen it being moved (Hare, Call & Tomasello, 2001). The "subject," the non-dominant chimp, could see what the dominant animal had or had not seen. In each case, he preferentially chose to head towards food for which the dominant animal had ignorance/a false belief about its location (Hare, Call & Tomasello, 2001). Thus, chimpanzees seem to be able to act on an implicit understanding of other animals' knowledge and ignorance, when testing conditions are ecologically valid (competition for food), and their behavior must be guided by tracking other animals' knowledge (Suddendorf & Whiten, 2003).

Like understanding desire, this stage of implicit belief understanding requires an understanding that others' mental states are private, internal, and can change independent of reality. The mentalism that emerges with the understanding of desire is thus further extended into an implicit understanding of belief (Wellman & Lagattuta, 2000). Even with evidence for implicit belief understanding, however, there is still no water-tight evidence that either 3-year-old children or apes can explicitly represent the *contents* of others' mental states. Indeed, the fact that children this age still fail perspective-taking tasks (even with controls for non-mentalizing task demands) is evidence that they cannot explicitly represent the content of others' mental states

(Mossler, Marvin & Greenberg, 1976; Liben, 1978; Flavell et al., 1981). To explicitly represent others' mental contents, another cognitive ability must emerge first.

Theory of Mind "Proper": Metarepresentation

Although 3-year-olds and chimpanzees can demonstrate implicit tracking of others' belief states, this does not mean that they understand the representational nature of beliefs. Knowledge and belief are referred to as "epistemic" mental states, as they are about knowledge representations and referents: agent –represents-> [proposition]. A statement about belief can be true whether or not the proposition that the belief represents is true. Understanding this representational nature of knowledge and belief means understanding the way that epistemic mental states refer to propositions about the world. Mentalism does not suffice for understanding representation. Rather, a new step in ToM development must occur, "metarepresentation," the ability to explicitly represent representations *as representations* (Perner, 1991; Leslie, 1994; Baron-Cohen, 1995). It is metarepresentation that enables children to pass explicit false belief tasks, and it is metarepresentation that apes lack (Suddendorf, 1999). The child can now understand that beliefs refer to propositions about the world, can explicitly represent the contents of those beliefs, and thus represent explicitly that beliefs can be mistaken. Passing an explicit false belief task is certain evidence of theory of mind capacities (Dennett, 1987).

However, the converse is not true. Many other cognitive abilities also contribute to being able to pass an explicit false belief task. Thus, if a person fails a false belief task, it does not necessarily mean that he/she lacks metarepresentation. It might be that he/she lacks one of the other cognitive abilities on which successful false belief task performance depends. In particular, solving false belief tasks depends on executive control, being able to inhibit the inappropriate response – what the subject knows to be the true state of reality – in order to answer with the

perhaps less salient correct response – what the other person’s mental state is (Carlson & Moses, 2001; Flynn, O’Malley & Wood, 2004; Carlson, Moses & Claxton, 2004). In fact, children can pass false belief tasks slightly earlier if the task demands are changed in such a way that not so much inhibitory control is required, for example, by making the current state of reality less salient (Wellman & Lagattuta, 2000). False belief tasks also depend on working memory and sequencing, as the subject has to keep in mind all the elements of the story as it unfolds in order, and how those elements are changing with respect to each other (Keenan, 1998; Stone, Baron-Cohen & Knight 1998). Thus, someone who has deficits in inhibiting a prepotent response or in working memory could easily fail a false belief task while having intact metarepresentational abilities.

Furthermore, metarepresentation may be just one example of a more general cognitive ability, embedding/recursion. To explicitly represent “X represents -> [proposition]” requires the ability to embed one proposition in another. If metapresentation is simply one type of recursion rather than a separate ability, difficulties with recursion could cause failures on the false belief task (Corballis, 2003).

There are many other cognitive tasks that use metarepresentation and recursion: complex syntax, self-representation, creativity, episodic memory and future planning (a.k.a. “mental time travel”), metamemory and counterfactual reasoning (Shimamura, Janowsky & Squire, 1990; Knight & Grabowecy, 1995; Suddendorf, 1999; Suddendorf & Fletcher-Flinn, 1999; De Villiers, 2000; Shimamura, 2000). Thus, recursion and metarepresentation may be general cognitive abilities, not limited to social cognition, that interact with mentalism to produce what we call explicit theory of mind. Indeed, there is evidence from neuroimaging and patient studies that understanding beliefs can be dissociated from metarepresentation and counterfactual reasoning (Saxe and Kanwisher, 2003; Samson, Apperly, Chiavarino & Humphreys, 2004). If

metarepresentation is found to be used by many other cognitive abilities besides ToM, then it would not be socially specific.

As an example of metarepresentation and recursion in another domain, De Villiers explains one way that development in syntactical abilities enables and precedes development in explicit belief representation. She argues that the ability to form embedded sentence complements, that is utterances of the form “agent –says-> subordinate clause,” e.g. “He said that he finished his peas,” or “She says that she saw the movie,” provides the representational structure needed for explicitly representing belief and knowledge (De Villiers, 2000). Sentences such as “Agent says that X,” however, are about observable things, utterances, rather than about private and changeable things, such as mental states. Thus, the metarepresentational ability that is needed to use and understand sentence complements is distinct from mentalism, from understanding the relationship between mental states and reality. In development, the ability to use and understand embedded sentences, both sentence complements and embedded relative clauses, precedes the ability to pass false belief tests (De Villiers & Pyers, 2002; Smith, Apperly & White, 2003). Though not a strict test of cause and effect, this suggests that a general metarepresentational capacity could be necessary before children can perform successfully on explicit false belief tasks.

The idea that explicit ToM is dependent on the metarepresentational competence needed for such complex grammatical structures is consistent with results on the cognitive abilities of chimps. Currently, chimps have been found to fail an explicit false belief task, indicating that they lack explicit metarepresentation (Call & Tomasello, 1999). They also do not show any recursive abilities: chimpanzees who have been taught to use signs and symbols to refer to things have never been observed to use complex syntax at all, much less any kind of syntactical embedding (Snowdon, 2001). Apes also do not show any evidence of either episodic memory or future

planning, also abilities that depend on metarepresentation (Suddendorf & Busby, 2003). Thus, metarepresentation and recursion seem to be uniquely human abilities.

The union of two abilities, an implicit understanding of the changeable nature of mental states and the ability to do metarepresentation of the explicit contents of those mental states results in having an explicit ToM in humans. Below, I discuss neuroscience research on ToM, and interpret the findings in terms of implicit mental state understanding (“mentalism”) and metarepresentation.

Neuroscience research on ToM: Metarepresentation ≠ ToM

Social neuroscience has been studying ToM for less than a decade, and thus neuroscience research on ToM is still very much in its infancy. We are only now beginning to learn from the methodological issues that developmental and comparative psychologists have had to work out over the past 30 years. Much ToM research in neuroscience has not been done with proper controls for working memory, inhibitory demands of tasks, or other executive functions, nor has it been done with a clear definition of which types of mental states (*e.g.*, intention, belief, desire) are being tapped by various tasks. The body of research in this area claims variously that ToM might be processed in superior temporal areas, temporal pole, the amygdala, temporal-parietal junction (TPJ), medial frontal cortex, orbitofrontal cortex (OFC), and/or frontal pole (Goel et al., 1995; Stone, Baron-Cohen & Knight, 1998; Gallagher et al., 2000; Fine, Lumsden & Blair, 2001; Happé et al., 2001; Stuss et al., 2001; Gallagher, Jack, Roepstorff & Frith, 2002; Gregory et al., 2002; Frith & Frith, 2003; Snowden et al., 2003; Stone et al., 2003; Grèzes, Frith & Passingham, 2004; Samson et al., 2004; Saxe, Carey & Kanwisher, 2004). Having just painted a picture of the complexity of ToM and the many cognitive abilities that may contribute to successful performance of ToM tasks, not to mention ToM developments after age 4, I believe it is not

surprising that the brain basis of ToM has not been narrowed down more. One reason that such a variety of brain areas have emerged as important for ToM in different research studies could be that these different areas may be subserving different aspects of ToM.

Given the review of ToM's components above, I believe the following four questions need to be answered before we will have a clear answer about ToM in the brain.

1. Do patients fail ToM tasks because of non-ToM task demands? Do any patients show deficits on a ToM task but not a control task that has the same executive function (EF) demands or verbal comprehension demands? Does changing the task to lessen the executive/comprehension demands improve patients' performance?

2. Are ToM and EF independent? Do patients' deficits in ToM correlate with deficits on EF measures tapping into relevant areas of EF: inhibitory control, working memory? Are there patients who perform highly on relevant EF measures while being impaired in ToM? Are there patients with EF deficits who can perform well on ToM tasks that minimize executive demands?

3. Does inferring belief require different brain systems as inferring other mental states, e.g., desire or intention, or are the same brain areas involved? Are there patients who perform poorly on measures tapping explicit metarepresentation of belief while still being able to perform well on tasks measuring an understanding of desire or intention, and vice versa?

4. Is metarepresentation/recursion separable from ToM? Are there patients who perform poorly on ToM measures while still being able to perform well on other tasks requiring metarepresentation and recursion, such as comprehension of embedded sentences, or passing a false photograph test?

Below, I discuss how social neuroscience has or has not provided answers to these questions in more detail. Each of the four questions above can also be addressed using neuroimaging, looking for commonalities and differences in areas activated by different kinds of tasks. I will focus primarily on patient research, as neuroimaging research is reviewed elsewhere

in this volume (Saxe, Chapter N; see also Saxe, Carey & Kanwisher, 2004), and as only patient research can answer questions about whether an area is crucial for a particular ability.

Questions 1 & 2. Do patients fail ToM tasks because of non-ToM task demands? Are ToM and EF independent? Science consists of finding evidence that is consistent with one hypothesis and inconsistent with alternative hypotheses. If a neurological patient is impaired on a ToM task, the obvious alternative hypothesis is that the patient failed the task because of task demands that have nothing to do with ToM, e.g., inhibitory control or working memory. The correct way to test ToM in patients is to use control conditions and comparison tasks to be able to rule out such an alternative hypothesis. These careful controls have sometimes been done, but not always.

One way to control for task demands is to vary the non-ToM task demands to see if this makes a difference in ToM performance. In patients who had lesions in left dorsolateral frontal cortex (DFC), Stone et al. (1998) found that when participants had to hold the elements of the story in working memory (which is the standard false belief task format), DFC patients often failed the false belief task (66% correct). However, they performed almost at ceiling on these same tasks when we removed the working memory load (98% correct), showing that their ToM metarepresentational capacities were intact. Clearly, patients can fail a false belief task because of non-ToM task demands.

In frontal patients, it is particularly important to do these types of controls. If no such controls have been done, then direct correlations between EF measures and ToM performance should be reported, as these can be informative. Stuss et al. (2001) found that patients with right orbitomedial or bifrontal lesions were impaired in tasks measuring perspective-taking and deception, but their tasks had strong working memory and inhibitory demands, as patients had to track a sequence of actions involving hiding an object. There were no control tasks with the same

demands but no ToM component. Because these patients were impaired on some EF measures as well, it is difficult to interpret the patients' deficits as truly reflecting deficits in ToM, particularly since no direct correlation between ToM performance and EF measures was reported. Snowden et al. (2003) report that frontotemporal dementia patients (who have orbitofrontal (OFC) damage) were impaired on making ToM inferences from eye gaze. On a test measuring whether patients would infer what someone else wanted from that person's eye gaze direction ("Which one does X want?"), patients would answer with the item representing what they wanted rather than what the stimulus person was looking at. Patients could have responded this way because of impulsivity and lack of inhibitory control rather than because of a failure in ToM *per se*. These patients were also impaired on EF measures, though again, no direct correlations between EF and ToM were reported.

Researchers working with frontal patients can learn from the example of Samson et al. (2004), who used an elegant control condition in their video false belief task with temporoparietal junction (TPJ) lesion patients. In the control tasks, memory and inhibitory demands were matched, but false belief attribution was not required. Thus, the TPJ patients' poor false belief task performance is more clearly attributable to deficits in ToM rather than non-ToM demands.

One can also control for non-ToM task demands by using different kinds of ToM tasks. Happé et al. (2001), Stone, Baron-Cohen & Knight (1998) and Gregory et al. (2002) got around some of the task demands in ToM tasks by assessing ToM in neurological patients without using false belief tasks, giving patients cartoons to interpret, or stories that required an understanding of, for example, desire, emotion, belief, deception, white lies and social faux pas (Happé's "Strange Stories Task," and Stone & Baron-Cohen's "Faux Pas Recognition Task"). Happé et al. (2001) found deficits on the ToM questions on these tasks in a patient with resection of medial frontal cortex. Stone, Baron-Cohen & Knight (1998) and Gregory et al. (2002) found deficits on

the Faux Pas task in patients with damage to orbitofrontal cortex. Stone et al. (2003) also found deficits on the Faux Pas task in two patients with bilateral amygdala lesions.

However, even in using ToM tasks that are not false belief tasks, it is still important to look at the relationship to EF. Happé et al. (2001) report that the medial frontal patient had severe EF deficits, particularly in inhibition and working memory tasks, so there remains some question as to why this patient failed ToM tasks. Gregory et al. (2002) generally found no correlation between Faux Pas task performance and some EF measures, but did find a relationship between perseverative errors on the Wisconsin and Faux Pas performance. This correlation may have been driven by a couple of patients whose errors on the Faux Pas task were perseverative, in which they kept giving the same answer in the same words. One patient with bilateral but primarily left amygdala damage in Stone et al. (2003) was impaired in making belief and intention judgments, but also had EF deficits. In these studies, it is thus difficult to conclude that ToM deficits are independent of EF deficits.

There are examples, however, of patients who demonstrate impaired ToM without impaired EF. One patient from Gregory et al. (2002) showed a striking dissociation between his ToM performance, which was poor, and his EF performance, which was close to ceiling (Gregory et al., 2002; Lough, Gregory & Hodges, 2002). Stone et al. (2002) also report a patient from the Stone et al. (1998) study with OFC, temporal pole and amygdala damage, who was impaired on a variety of theory of mind tasks and had intact executive function. The patient was further impaired in the ability to tell whether someone might be cheating another person, which could possibly tap into theory of mind, but performed normally on a control task matched exactly for executive and non-executive task demands (Stone et al., 2002). Fine, Lumsden & Blair (2001) report a patient with left amygdala damage acquired in childhood who had high scores on EF tasks, particularly inhibition tasks, but was severely impaired on a variety of ToM tests, including

false belief tasks. Thus, his poor performance on false belief tasks clearly cannot be accounted for by difficulties with inhibition. Stone et al. (2003) report a patient with bilateral amygdala damage acquired in adulthood, with damage primarily on the right, who was impaired in the Faux Pas Recognition task and in Reading the Mind in the Eyes, and unimpaired on executive function. In looking for ToM deficits independent of EF deficits, patient research points to the amygdala, orbitofrontal cortex (OFC) and temporo-parietal junction (TPJ) as possible key areas. Medial frontal cortex may also be involved, but results on independence from EF are inconclusive.

Question 3: Does inferring belief require different brain systems as inferring other mental states, e.g., desire or intention, or are the same brain areas involved? Many ToM researchers in neuroscience have used tasks that measure multiple types of inferences, including epistemic inferences about belief, and inferences about desire or intention (Happé's "Strange Stories Task," and Stone & Baron-Cohen's "Faux Pas Recognition Task", tasks used with fMRI in Saxe & Kanwisher, 2003). Attributing intentions is an early building block of ToM, something that great apes and very young children can do. Attributing desires taps into mentalism, but does not require metarepresentation. Thus, it is important to try to tease apart whether all of these abilities are using the same neural substrates, or whether different kinds of mentalistic inferences depend on different brain areas.

We can answer these questions if we use more fine-grained methods. Researchers can give patients or participants in a scanner multiple tasks, some that require only belief, and others that require desire or intention, and report the results for belief, desire and intention separately. In Stone et al. (1998), patients with orbitofrontal (OFC) damage were impaired on the Recognition of Faux Pas task, but performed at ceiling on false belief tasks. Many of the OFC patients' errors on the Faux Pas task in both Stone et al. (1998) and Gregory et al. (2002) were reflected in statements such as, "Well, he meant to put him down," or "He wanted to make her

feel bad so he could feel like the big man.” They seemed to be errors in judging whether or not the faux pas was committed accidentally or intentionally. In Gregory et al. (2002), only the most severely affected fronto-temporal dementia patients, whose damage may have spread beyond OFC, had difficulty with false belief tasks.

Further support for the idea that OFC is not involved in any kind of metarepresentational ToM inference comes from results with the same OFC patients who were tested in Stone et al. (1998) on a test (the Soap Opera Task) measuring ability to make 0-, first-, second-, and third-order mental state inferences. Participants read fairly complex stories about topics such as spies, embezzling, or extramarital affairs, and then were asked to make true/false judgments about statements involving mental state inferences and control statements about details of the stories. Statements about mental states requiring no metarepresentation (0-order) were all about character’s likes or desires², e.g., “Tim fancies Maria,” or “The children like Easter candy.” Third-order belief statements required the highest level of recursion, e.g., “John thought that Sue believed that Mary thought that X.” Control statements were matched for grammatical complexity with ToM questions, and had equal levels of embedded clauses. Both mental state and control statements were constructed so that all the levels of grammatical embedding had to be parsed to get the correct answer; the participant could not simply make the correct choice based on one clause. OFC patients and controls performed equally on this task, scoring highly on all questions. Both groups made more errors on the 2nd- and 3rd-order statements, but made no more errors on ToM than on non-ToM statements. Thus, with a more difficult task tapping metarepresentation in both ToM and non-ToM linguistic statements, OFC patients showed no deficits. These results make it unlikely that OFC is involved in any metarepresentational aspects of ToM.

The picture of OFC's role in ToM is complicated by neuroimaging results concerning belief. Most neuroimaging does not find OFC activation specific to belief tasks, though it is difficult to get a good signal from OFC in fMRI. However, a recent nonverbal belief task did find OFC activation specific to watching someone perform an action with a false expectation vs. a true expectation (Grèzes, Frith & Passingham, 2004). Since the task looked at expectations, rather than belief statements with content, it is possible that the OFC involvement could reflect judgments of intended vs. unintended actions. Liu, Sabbagh, Gehring & Wellman (2004) used the false belief task, with a few true belief catch trials, and looked at ERP components that are closely time-locked to the point at which participants make a belief judgment. New statistical techniques allow some rough localization of where the signal generator for a component is. They report that the ERP component specific to the belief questions and not control questions is statistically inconsistent with a signal generator in dorsal or medial frontal cortex, but is consistent with a generator in left OFC. However, it might be difficult to rule out a generator close to OFC, such as temporal pole. It would also be important to see the results for true and false belief items separately, to rule out the possibility that the signal results from inhibition required when answering about false belief. It may take future research with a technology such as MEG, with better spatial *and* temporal resolution, and using both verbal and nonverbal ToM tasks, to clarify the meaning of these studies. I believe the most parsimonious interpretation of all results on ToM and OFC may be that OFC is mediating judgments about intentional actions, but not belief. OFC is considered an evolutionarily older part of the frontal lobes, and thus it makes some sense that it would handle judgments of intentional behaviour rather than computing meta-representational mental state inferences.

Unfortunately, it has been difficult to separate belief and desire in both patient and neuroimaging results. Happé et al. (2001) did not report separate results for the stories in the

Strange Stories task or the cartoons that assess belief attribution versus an understanding of desire, so it is unclear if medial frontal cortex is involved in belief, desire, or both. If it were specific to representing belief, then it should be active for both true belief and false belief, yet it is not as active during true belief as false belief attribution (Fletcher et al., 1995; Saxe, Kanwisher & Carey, 2004). Fine, Lumsden & Blair (2001) used false belief tasks, and the same tasks used by Happé et al. (2001) with their patient with amygdala damage. He was impaired on all these tasks, but again, on the tasks that asked about both belief and desire, separate results for belief inferences and other mental state inferences are not reported. With both the medial frontal and amygdala patients, it is possible that their understanding of desire was completely unimpaired while their understanding of belief was impaired. Samson et al. (2004) report only results for location change false belief tasks, thus we have no information about TPJ patients' understanding of desire as a mental state. Saxe and Kanwisher (2003) report that the TPJ is more active during stories requiring desire inferences than physical inferences, as is true for belief compared to physical inferences. Saxe, Carey & Kanwisher (2004) note that the TPJ is more active for belief than desire inferences, but also that the desire stories may also have elicited belief attributions. In future research reports on patients, I believe it is important to report results separately for desire and belief, and in neuroimaging to use tasks that cleanly assess desire and belief separately. There are also a range of mental states related to desire, e.g. adoration, disgust, and so a pure test of belief vs. desire would ideally sample all of these mental states. If areas active for belief tasks always turn out to also be active for understanding desire, then this would point to these areas being involved in mentalism more generally, rather than just in belief.

Question 4: Is metarepresentation/recursion separable from ToM? Developmental and evolutionary considerations point to mentalism, i.e., an understanding of the nature of mental things, being separable from metarepresentation/recursion required to do false belief tasks. If so,

then these abilities might be dissociable in patient or neuroimaging research. There is no published patient research that has addressed the question of whether areas involved in ToM might be involved in other non-ToM tasks that require metarepresentation and recursion. Saxe and Kanwisher (2003) did look at separability of ToM and metarepresentation in fMRI. The false photograph task requires an inference about whether or not a photograph, a non-mental representation, has changed if the state of reality changes after the picture is taken. Thus it requires metarepresentation, representing the representational nature of the photograph, but not mentalism. Participants were scanned while doing false belief tasks and false photograph tasks. Activation in the TPJ, superior temporal pole and medial portions of frontal pole was significantly greater during the belief tasks. However, they do not report whether these areas were more active during the false photograph stories than during stories about physical descriptions of people and objects that do not involve metapresentation. If these areas were differentially active during the false photograph test, then that would be some evidence for their involvement in metarepresentation. Clearly, TPJ, medial portions of frontal pole and temporal pole seem to be involved in mentalism, and possibly, this ability can be separated from metarepresentation.

With respect to TPJ in particular, some further research distinguishing it from areas involved in recursion would be helpful. Some regions very close to the parts of TPJ that were reported for ToM in Saxe & Kanwisher (2003) have also been found to be active during a specific kind of grammatical task that requires recursion, the processing of embedded relative clauses (frontal areas were also active; Caplan et al., 2001; Cooke et al., 2001). These areas do not completely overlap with the TPJ areas for ToM, but some direct tests would be useful. Using a test such as the Soap Opera task that I used with OFC patients would directly compare embedding/recursion in both ToM and non-ToM control questions. Using this test with patients with medial frontal damage, TPJ damage (provided they were not aphasic) or frontal pole

damage, all areas thought to be involved in belief representation, could help uncover dissociations between recursion and mentalism in ToM inferences. A study directly comparing the processing of embedded sentences and ToM in fMRI would solidify the conclusion that TPJ is involved in mentalism rather than recursion and metarepresentation.

Neuroimaging research with implicit belief tasks would also be important. The first study ever to test ToM with fMRI used a task that required only implicit inferences about beliefs, which represents a true advantage over many imaging studies (Goel et al., 1995). Most imaging studies have looked at explicit, deliberate reasoning about ToM, often including deliberate instructions to think about character's motivations or mental states (Liu et al., 2004; Saxe, Carey & Kanwisher, 2004). ToM inferences in everyday life are made on the fly, implicitly. Tasks requiring deliberate inferences may not tap into these processes in the same way. Implicit belief attribution tasks, perhaps styled after this first one used, or after Perner & Garnham's (2001) implicit false belief task (without the running around with mattress pads), would be valuable additions to imaging research in this area, as they might help identify areas that are involved in mentalism but not metarepresentation of belief.

The maturation of ToM research in neuroscience

Decades of research on ToM in developmental psychology and primatology have given us a detailed picture of its precursors and components. Developmental research into why children have difficulty with false belief tasks, in particular, has provided insight into how people can seem to be impaired on certain ToM tasks because of limitations in non-ToM cognitive abilities. Neuroscience research on ToM is just beginning to take these methodological lessons into account. We are also just beginning to make distinctions between ToM and other related abilities such as recursion.

I believe the mentalism evident in children's and primate's understanding of desire and their implicit understanding of belief should be carefully distinguished from metarepresentation of belief and recursion, because they emerge at different points in development and evolution. Though further research will solidify this conclusion, medial portions of frontal pole, TPJ and superior temporal pole are probably being activated in ToM tasks because they subserve mentalism, rather than metarepresentation and recursion in general. I suggest that these areas mediate what is specifically social about theory of mind, the mentalism required to understand that belief states and desires can change independent of reality, and form the neural basis of implicit mental state understanding. As such, the computations carried out by these areas would be maturing between 24 – 40 months of age in human children, and could be those shared with other great apes. To the extent that brain regions are involved in *specifically social* computations, they may not be involved in *uniquely human* computations.

Metarepresentation, recursion and executive control are not at all limited to ToM. They enable language, complex tool manufacture, future planning, episodic memory, and explicit cultural transmission of knowledge, all things that are the hallmarks of the cognitive uniqueness of *Homo sapiens sapiens* (Suddendorf, 1999; Corballis, 2003). Uniquely human abilities are likely to be frontal and temporal, as these areas are clearly disproportionately larger in humans compared to other recent hominids and great apes, more so than other cortical regions (Semendeferi et al., 2001; Lieberman et al., 2002). The degree of executive control over cognition in humans is unique among primates. This at least we know is mediated by the frontal lobes. In contrast, the brain areas involved in metarepresentation and recursion remain unknown. Uniquely human aspects of language depend on recursion, and such complex syntactic abilities seem to involve areas in left temporal-parietal and inferior frontal regions (Caplan et al., 2001; Cooke et al., 2001).

Overall, areas involved in the social aspects of ToM seem well-positioned, anatomically, to interact with areas involved in executive function and recursion. Further research using both patient data and neuroimaging, testing a variety of both ToM and non-ToM tasks that require executive control, metarepresentation and recursion can clarify these issues. For now, we can understand the processing of ToM in the brain as the interaction of several regions, some specifically social, some not. The operation of human social intelligence in the brain involves areas and functions that are shared with other social mammals, such as gaze direction detection in the STS, some shared only with primates, such as gaze following, some shared only with great apes, such as joint attention or detecting intentionality, and some that are uniquely human. As neuroscience research on ToM matures, and takes these complexities into account, we will have a clearer picture of how the brain areas involved in ToM might have evolved and how different components of ToM interact with each other.

Footnotes

¹ In an unexpected contents task, the subject is shown a container that is clearly labeled as if it contains one kind of thing; for example, a candy box clearly indicates that it contains candy. The subject is shown that it really contains a quite different thing, e.g. there are pencils inside the candy box. Then the subject is asked what another person, who hasn't seen what's inside the box, will think is in there. Control questions usually ask about what was true originally, what the subject thought originally, and what is true right now.

² Some would describe desire statements as involved "first-order intentionality" (Dennett, 1987). When I use 0-order, 1st-order, 2nd-order and 3rd-order here, I am referring to the level of grammatical embedding.

References

- Baron-Cohen, S. (1995). *Mindblindness: An Essay on Autism and Theory of Mind*. Cambridge, MA: MIT Press.
- Baron-Cohen, S., Baldwin, D. & Crowson, M. (1997). Do children with autism use the speaker's direction of gaze strategy to crack the code of language? *Child Development*, *68*(1), 48-57.
- Brothers, L. (1997). *Friday's Footprint: How Society Shapes the Human Mind*. Oxford: Oxford University Press.
- Call, J. & Tomasello, M. (1998). Distinguishing intentional from accidental actions in orangutans (*Pongo pygmaeus*), chimpanzees (*Pan troglodytes*) and human children (*Homo sapiens*). *Journal of Comparative Psychology*, *112*(2), 192-206.
- Call, J. & Tomasello, M. (1999). A nonverbal false belief task: the performance of children and great apes. *Child Development*, *70*(2), 381-95.
- Caplan, D., Vijayan, S., Kuperberg, G., West, C., Waters, G., Greve, D.D. & Anders, M. (2002). Vascular responses to syntactic processing: Event-related fMRI study of relative clauses. *Human Brain Mapping*, *15*(1), 26-38.
- Carlson, S. & Moses, L. (2001). Individual differences in inhibitory control and children's theory of mind. *Child Development*, *72*(4), 1032-1053.
- Carlson, S., Moses, L. & Klaxton, L. (2004). Individual differences in executive functioning and theory of mind: An investigation of inhibitory control and planning ability. *Journal of Experimental Child Psychology*, *87*(4), 299-319.
- Cooke, A., Zurif, E.B., DeVita, C., Alsop, D., Koenig, P., Detre, J., Gee, J., Pinango, M., Balogh, J. & Grossman, M. (2001). Neural basis for sentence comprehension: grammatical and short-term memory components. *Human Brain Mapping*, *15*, 80-94.
- Corballis, M. (2003). Recursion as the key to the human mind. In K. Sterelny and J. Fitness (Eds.), *From mating to mentality: Evaluating evolutionary psychology* (pp. 155-171). New York, Psychology Press.
- Csibra, G., Biro, S., Koos, O., Gergely, G. (2003). One-year-old infants use teleological representations of actions productively. *Cognitive Science*, *27*(1), 111-133.
- Dennett, D. (1987). *The Intentional Stance*. Cambridge, MA: MIT Press.

De Villiers, J. (2000). Language and theory of mind: What are the developmental relationships? In Baron-Cohen, S., Tager-Flusberg, H. & Cohen, D. (Eds.), *Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience, 2nd Edition* (pp. 83-123). Oxford: Oxford University Press.

De Villiers, J. & Pyers, J. (2002). Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief understanding. *Cognitive Development, 17*, 1037-1060.

Fine, C., Lumsden, J. & Blair, J. (2001). Dissociation between 'theory of mind' and executive functions in a patient with early left amygdala damage. *Brain, 124*, 287-298.

Flavell, J., Everett, Croft & Flavell, E. (1981). Young children's knowledge about visual perception—Further evidence for the Level 1-Level 2 distinction. *Developmental Psychology, 17*, 99-103.

Fletcher, P., Happé, F., Frith, U., Baker, S., Dolan, R.J., Frackowiak, R. & Frith, C. (1995). Other minds in the brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition, 57*(2), 109-128.

Flynn, E., O'Malley, C. & Wood, D. (2004). A longitudinal, microgenetic study of the emergence of false belief understanding and inhibition skills. *Developmental Science, 7*(1), 103-115.

Foley, R. (1997). *Humans before humanity*. London: Blackwell.

Franco, F. & Butterworth, G. (1996). Pointing and social awareness: Declaring and requesting in the second year. *Journal of Child Language, 23*(2), 307-336.

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

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

Gallagher, H.L., Happe, F., Brunswick, N. Fletcher, P., Frith, U., Frith, C.D. (2000). Reading the mind in cartoons and stories: an fMRI study of 'theory of the mind' in verbal and nonverbal tasks. *Neuropsychologia, 38*(1), 11-21.

Gallagher, H.L., Jack, A.I., Roepstorff, A. & Frith C.D. (2002). Imaging the intentional stance in a competitive game. *Neuroimage, 16*, 814-821.

Gibbons, A. (2002). In search of the first hominids. *Science, 295*, 1214-1219.

Goel, V., Grafman, J., Sadato, N. & Hallett, M. (1995). Modeling other minds. *Neuroreport, 6*(13), 1741-1746.

Gregory, C. , Lough, S., Stone, V.E., Erzinclioglu, S., Martin, L., Baron-Cohen, S. & Hodges, J. (2002). Theory of mind in frontotemporal dementia and Alzheimer's disease: Theoretical and practical implications. *Brain*, 125, 752-764.

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

Happé, F., Mahli, G.S. & Checkley, S. (2001). Acquired mind-blindness following frontal lobe surgery. A single case study of impaired 'theory of mind' in a patient treated with stereotactic anterior capsulotomy. *Neuropsychologia*, 39, 83-90.

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

Hare, B., Call, J. & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behaviour*, 61, 139-151.

Jellema, T., Baker, C.I., Wicker, B. & Perrett, D.I. (2000). Neural representation for the perception of the intentionality of actions. *Brain & Cognition. Special Issue: Cognitive neuroscience of actions*, 44(2), 280-302 .

Keenan, T. (1998). Memory span as a predictor of false belief understanding. *New Zealand Journal of Psychology*, 27(2), 36-43.

Knight, R.T. & Grabowecy, M. (1995). Escape from linear time: Prefrontal cortex and conscious experience. In M.S. Gazzaniga (Ed.), *The Cognitive Neurosciences*. Cambridge, MA: MIT Press.

Kumashiro, M., Ishibashi, H., Itakura, S. & Iriki, A. (2002). Bidirectional communication between a Japanese monkey and a human through eye gaze and pointing. *Current Psychology of Cognition*, 21(1), 3-32.

Kumashiro, M., Ishibashi, H., Uchiyama, Y., Itakura, S., Murata, A. & Iriki, A. (2003). Natural imitation induced by joint attention in Japanese monkeys. *International Journal of Psychophysiology*, 50, 81-99.

Leavens, D.A., Hopkins, W.D. & Thomas, R.K. (2004). Referential communication by chimpanzees (*Pan troglodytes*). *Journal of Comparative Psychology*, 118(1), 48-57.

Leslie, A.M. (1987). Pretence and representation: The origins of 'theory of mind.' *Psychological Review*, 94, 412-426.

Leslie, A.M. (1994). Pretending and believing: Issues in the theory of ToMM. *Cognition*, 50, 211-238.

- Leslie, A.M. & Frith, U. (1990). Prospects for a cognitive neuropsychology of autism: Hobson's choice. *Psychological Review*, *97*(1), 122-131.
- Liben, L.S. (1978). Perspective-taking skills in young children: Seeing the world through rose-colored glasses. *Developmental Psychology*, *14*(1), 87-92.
- Lieberman, D.E., McBratney, B.M. & Krovitz, G. (2002). The evolution and development of cranial form in *Homo sapiens*. *Proceedings of the National Academy of Science, USA*, *99*(3), 1134-1139.
- Lillard, A.S. Zeljo, A., Curenton, S., Kaugars, A.S. (2000). Children's understanding of the animacy constraint on pretense. *Merrill-Palmer Quarterly*, *46*(1), 21-44.
- Liu, D., Sabbagh, M.A., Gehring, W.J. & Wellman, H. (2004). Decoupling beliefs from reality in the brain: an ERP study of theory of mind. *Neuroreport*, *15*(6), 991-995.
- Lough, S., Gregory, C. & Hodges, J. (2002). Dissociation of social cognition and executive function in frontal variant frontotemporal dementia. *Neurocase. Special Issue: Frontotemporal dementia: Part II*, *7*(2), 123-130.
- McGrew, (2001). The nature of culture. In F. deWaal (Ed.), *Tree of Origin: What Primate Behavior Can Tell Us about Human Social Evolution* (pp. 231-254). Cambridge, MA: Harvard University Press.
- Mossler, D.G., Marvin, R.S. & Greenberg, M.T. (1976). Conceptual perspective taking in 2- to 6-year-old children. *Developmental Psychology*, *12*(1), 85-86.
- Perner, J. & Garnham, W.A. (2001). Actions really do speak louder than words--But only implicitly: Young children's understanding of false belief in action. *British Journal of Developmental Psychology*, *19*(3), 413-432.
- Povinelli, D.J. & O'Neill (2000). In Baron-Cohen, S., Tager-Flusberg, H. & Cohen, D. (Eds.), *Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience, 2nd Edition* (pp.). Oxford: Oxford University Press.
- Povinelli, D.J., Theall, L.A., Reaux, J.E., Dunphy-Lelii S. (2003). Chimpanzees spontaneously alter the location of their gestures to match the attentional orientation of others. *Animal Behaviour*, *66*(1), 71-79.
- Pratt, C. & Bryant, P. (1990). Young children understand that looking leads to knowing (so long as they are looking into a single barrel). *Child Development*, *61*(4), 973-982.
- Repacholi, B.M. & Gopnik, A. (1997). Early reasoning about desires: Evidence from 14- and 18-month-olds. *Developmental Psychology*, *33*(1), 12-21.

Samson, D., Apperly, I.A., Chiavarino, C. & Humphreys, G.W. (2004). The left temporoparietal junction is necessary for representing someone else's belief. *Nature Neuroscience*, 7(5), 499-500.

Saxe, R. & Kanwisher, N. (2003). People thinking about people: the role of the temporoparietal junction in "theory of mind." *Neuroimage*, 19, 1835-1842.

Saxe, R., Carey, S. & Kanwisher, N. (2004). Understanding other minds: Linking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, 55, 87-124.

Semendeferi, K., Armstrong, E., Schleicher, A., Zilles, K. & Van Hoesen, G.W. (2001). Prefrontal cortex in humans and apes: a comparative study of area 10. *American Journal of Physical Anthropology*, 114(3), 224-41.

Shimamura, A.P. (2000). Toward a cognitive neuroscience of metacognition. *Consciousness & Cognition*, 9, 313-23

Shimamura, A.P., Janowsky, J.S. & Squire, L.R. (1990). Memory for the temporal order of events in patients with frontal lobe lesions and amnesic patients. *Neuropsychologia*, 28(8), 803-813.

Smith, M., Apperly, I., White, V. (2003). False belief reasoning and the acquisition of relative clause sentences. *Child Development*, 74(6), 1709-1719.

Snowden, J.S., Gibbons, Z., Blackshaw, A., Doubleday, E., Thompson, J., Craufurd, D., Foster, J., Happé, F. & Neary, D. (2003). Social cognition in frontotemporal dementia and Huntington's disease. *Neuropsychologia*, 41, 688-701.

Snowdon, C.T. (2001). From primate communication to human language. In F. deWaal (Ed.), *Tree of Origin: What Primate Behavior Can Tell Us about Human Social Evolution* (pp. 195-227). Cambridge, MA: Harvard University Press.

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

Stone, V.E., Baron-Cohen, S., Calder, A.C., Keane, J. & Young, A.W. (2003). Acquired theory of mind impairments in individuals with bilateral amygdala lesions. *Neuropsychologia*, 41, 209-220.

Stone, V.E., Cosmides, L., Tooby, J., Kroll, N. & Knight, R.T. (2002). Selective impairment of reasoning about social exchange in a patient with bilateral limbic system damage. *Proceedings of the National Academy of Sciences, USA*, 99(17), 11531-11536.

Stuss, D.T., Gallup, G. & Alexander, M. (2001). The frontal lobes are necessary for "theory of mind." *Brain*, 124(2), 279-286.

Suddendorf, T. (1999). The rise of the metamind. In M.C. Corballis & S. Lea (Eds.), *The descent of mind: Psychological perspectives on hominid evolution* (pp. 218-260). London: Oxford University Press.

Suddendorf, T. (in press). How primatology can inform us about the human mind. *Australian Psychologist*.

Suddendorf, T. & Busby, J. (2003). Mental time travel in animals? *Trends in Cognitive Sciences*, 7, 391-396.

Suddendorf, T., & Fletcher-Flinn, C.M. (1999). Children's divergent thinking improves when they understand false beliefs. *Creativity Research Journal, Special Issue: Longitudinal Studies of Creativity*, 12, 115-128.

Suddendorf, T., & Whiten, A. (2001). Mental evolution and development: evidence for secondary representation in children, great apes and other animals. *Psychological Bulletin*, 127(5), 629-650.

Suddendorf, T., & A. Whiten (2003). Reinterpreting the Mentality of Apes. In K. Sterelny and J. Fitness (Eds.), *From mating to mentality: Evaluating evolutionary psychology* (pp. 173-196). New York, Psychology Press.

Tooby, J. & Cosmides, L. (1992). The psychological foundations of culture. In J. Barkow, L. Cosmides & J. Tooby (Eds.), *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. (pp. 19-136). New York: Oxford University Press.

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

Wellman, H. & Lagattuta, K.H. (2000). Developing understandings of mind. In Baron-Cohen, S., Tager-Flusberg, H. & Cohen, D. (Eds.), *Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience, 2nd Edition* (pp. 21-49). Oxford: Oxford University Press.

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

Wellman, H. & Liu, D. (2004). Scaling Theory of Mind tasks. *Child Development*, 75(2), 523-541.

Wellman, H. & Wooley, J. (1990). From simple desires to ordinary beliefs: The early development of everyday psychology. *Cognition*, 35(3), 245-275.

Wildman, D.E., Uddin, M., Liu, G., Grossman, L.I. & Goodman, M. (2003). Implications of natural selection in shaping 99.4% nonsynonymous DNA identity between humans and chimpanzees: Enlarging genus *Homo*. *Proceedings of the National Academy of Sciences, USA*, 100(12), 7181-7188.

Woodward, A. (1999). Infants' ability to distinguish between purposeful and non-purposeful behaviors. *Infant Behavior & Development*, 22(2), 145-160.

Yoder, A.D., Cartmill, M., Ruvolo, M., Smith, K. & Vilgalys, R. (1996). Ancient single origin for Malagasy primates. *Proceedings of the National Academy of Science, USA*, 93(10), 5122-5126.

Figure 1. Our primate family tree, showing branching points leading to monkeys, great apes, hominids and humans. Time is measured in mya = millions of years ago, except where marked kya = thousands of years ago. Information on when certain branchings occurred is based on genetic analyses and has been compiled from Yoder et al. (1996), Foley (1997), Gibbons (2000) and Wildman et al. (2003).