

- Shelton, A. L., & McNamara, T. P. (2001b). Visual memories from nonvisual experiences. *Psychological Science*, 12(4), 343–347.
- Simons, D. J., & Rensink, R. A. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences*, 9(1), 16–20.
- Stenning, K. (2002). *Seeing reason, image and language in learning to think*. Oxford, UK: Oxford University Press.
- Stenning, K., & Oberlander, J. (1995). A cognitive theory of graphical and linguistic reasoning: Logic and implementation. *Cognitive Science*, 19, 97–140.
- Strawson, P. F. (1959). *Individuals: An essay in descriptive metaphysics*. London: Methuen.
- Strawson, P. F. (1974). *Subject and predicate in logic and grammar*. London: Methuen.
- Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1996). Using eye movements to study spoken language comprehension: Evidence for visually mediated incremental interpretation. In T. Inui & J. L. McClelland (Eds.), *Attention and performance XVI* (pp. 457–478). Cambridge, MA: MIT Press.
- Todd, P. M., & Gigerenzer, G. (2000). Précis of Simple heuristics that make us smart. *Behavioral and Brain Sciences*, 23, 727–780.
- Tomasello, M., Carpenter, M., & Liszkowski, U. (2007). A new look at infant pointing. *Child Development*, 78(3), 705–722.
- Treisman, A. (1992). Perceiving and re-perceiving objects. *American Psychologist*, 47, 862–875.
- Triesch, J., Ballard, D. H., Hayhoe, M. M., & Sullivan, B. T. (2003). What you see is what you need. *Journal of Vision*, 3, 86–94.
- Ullman, S. (1984). Visual routines. *Cognition*, 18, 97–159.
- Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, 21, 615–665.
- Van Oeffelen, M., & Vos, P. (1982). Configurational effects on the enumeration of dots: Counting by groups. *Memory and Cognition*, 10, 396–404.
- Verblunsky, S. (1951). On the shortest path through a number of points. *Proceedings of the American Mathematical Society*, 2(6), 904–913.
- Yantis, S. (1992). Multielement visual tracking: Attention and perceptual organization. *Cognitive Psychology*, 24, 295–340.
- Zimmer, H. D., Cohen, R. L., Guynn, M. J., Engelkamp, J., Kormi-Nouri, R., & Foley, M. A. (2001). *Memory for action: A distinct form of episodic memory?* Oxford, UK: Oxford University Press.

## An Evolutionary Cognitive Neuroscience Perspective on Human Self-awareness and Theory of Mind

Farah Focquaert, Johan Braeckman and Steven M. Platek

*The evolutionary claim that the function of self-awareness lies, at least in part, in the benefits of theory of mind (TOM) regained attention in light of current findings in cognitive neuroscience, including mirror neuron research. Although certain non-human primates most likely possess mirror self-recognition skills, we claim that they lack the introspective abilities that are crucial for human-like TOM. Primate research on TOM skills such as emotional recognition, seeing versus knowing and ignorance versus knowing are discussed. Based upon current findings in cognitive neuroscience, we provide evidence in favor of an introspection-based simulation theory account of human mindreading.*

**Keywords:** Self-Awareness; Simulation Theory; Theory of Mind

### 1. Introduction

There is an ongoing debate as to whether or not non-human primates, particularly chimpanzees, display theory of mind (TOM) dependent skills, such as the display of empathy and the detection of deception. TOM, or 'mindreading', can be defined as the attribution of mental states, thoughts and emotions to an individual, in order to explain and predict its behavior (Premack & Woodruff, 1978). Consider empathy: Several studies documenting the consoling behavior of chimpanzees (see de Waal, Thompson & Proctor, 2005) and numerous field observations by primate researchers (of, for example, the care a mother chimpanzee displays towards her offspring; see Goodall, 1990), putatively point to the presence of certain empathic behaviors in at least some non-human great apes. Nevertheless, except for consoling behavior, little systematic research has been done specifically on empathy in non-human primates (de Waal et al., 2005), and it remains unclear how to interpret the available data.

Most researchers agree that a human-like TOM system is not to be found in any non-human primate species. Nevertheless, it would be inaccurate to claim that all non-human primate species lack human-like TOM-dependent skills altogether. On the contrary, certain non-human primates such as chimpanzees can be said to possess at least some aspects, albeit basic ones, of a human-like TOM system. The fact that some non-human primates utilize tactical deception, grasp the difference between seeing and knowing, and the difference between ignorance and knowing (Byrne & Whiten, 1990; Gómez, 1998, 2004; Hare, Call, Agnetta & Tomasello, 2000; Hare, Call & Tomasello, 2001), for example, shows us that non-human great ape species have some *understanding* of the causes of others' behavior. In this paper we will ask whether or not human mindreading skills differ from non-human mindreading skills, and, if so, how this might be explained based upon current scientific research in a variety of fields such as experimental primatology, philosophy of mind, developmental psychology, cognitive neuroscience and genetics.

## 2. Do Non-human Animals Possess Human Mindreading Skills?

Recently, Parr (2003a) argued that her work on emotional recognition reveals that chimpanzees might possess a basic kind of emotional awareness that humans and chimpanzees can be said to have in common. In a 'matching to meaning' (MTM) task chimpanzees correctly related short emotional video scenes portraying either negative or positive emotional valence—for example, a scene named 'inject' portrayed chimpanzees being injected with darts and needles—to pictures portraying either negative or positive emotional facial expressions, such as scream faces and bared-teeth displays (Parr, 2001). According to Parr (2003a), the chimpanzees can not rely on perceptual similarities to solve this task, but instead need to have an understanding of the shared emotional meaning depicted by the video scene on the one hand, and the correct facial expression on the other hand. Such perceptual similarities can play a role in 'matching to sample' (MTS) tasks. An MTS task consists, for example, in matching two out of three pictures of unfamiliar others. Two of these pictures represent two different animals with the same facial expression, such as a scream face, and a third picture represents yet another animal with a neutral facial expression. In this task, the chimpanzees are asked to match stimuli based on their physical similarity, e.g., matching facial expressions, whereas in the MTM task chimpanzees are required to match stimuli based on their *underlying* emotional similarity (Parr, 2003b).

Unlike monkeys, chimpanzees are able to recognize emotional similarity between stimuli. Nevertheless, the data (Parr, 2001) do not show that chimpanzees have a conscious understanding of the relationship between the emotion and the facial expression. Parr argues (2003b) that the most reasonable explanation may be that—by virtue of having experience with the situations that are depicted—the chimpanzees' responses are the result of emotional contagion. The observation of a stressful situation, such as a chimp being injected with a dart, elicits a negative feeling

in the chimpanzee and urges the chimpanzee to choose the facial expression related to this feeling. According to Parr (2003a), the chimpanzees' ability to make discriminations based on emotional information nevertheless endows them with a basic kind of emotional awareness that is a likely *precursor* to more cognitive forms of mindreading found in humans.

It is unclear if the emotional awareness demonstrated by chimpanzees is the relevant kind of emotional awareness, and if the capacities demonstrated by chimpanzee involve the attribution of mental states or not. According to Vonk and Povinelli (2006), these facts can't be derived from the Parr's (2001) experiment. They argue that the chimpanzees' success at this task might be the result of correctly associating the behavior of the chimpanzee in the scene with the appropriate facial expression, without representing the underlying emotion in any way.

Vonk and Povinelli (2006) argue that the emotions in the videos cannot be depicted in the absence of the observable behavior with which they are correlated. For example, when viewing the 'inject' scene it is possible that the chimpanzee performing the task does not represent the underlying emotion *fear* in any way, but instead correctly associates the typical behavior of a chimpanzee in fear with the appropriate facial expression. They conclude that there is no need to invoke any conceptual understanding or recognition of the underlying emotions to explain the obtained results. There need not be any human-like TOM skills involved.

According to Vonk and Povinelli's (2006) 'Unobservability Hypothesis' one of the crucial differences between human and non-human minds is our ability to think and reason about *unobservable entities*, such as God, ghosts and other minds. Although many species form concepts about observable things and are able to use these concepts in very flexible ways, only humans form concepts about purely hypothetical things. In their view, a chimpanzee's mind never loses touch with the observable world around it. Their hypothesis has serious implications as to how a chimpanzee's TOM system might differ from our own. A human-like TOM system entails that, aside from reasoning about behavior, one additionally reasons about the underlying mental states, or *unobservables*, that give rise to that behavior.

In an attempt to determine whether chimpanzees can reason about mental states, the Povinelli lab set out to assess chimpanzees' understanding of the mental state 'knowing'. When initial attempts failed to clearly establish this—for example chimpanzees do not appear to understand that 'seeing' something hidden leads to 'knowing' its location—they focused their research on the aspect 'seeing' itself. Although chimpanzees appear to be able to follow the gaze of others (Povinelli & Eddy, 1996) and exploit gaze cues to locate hidden food (Povinelli, Bierschwale & Cech 1999; Povinelli, Dunphy-Lelii, Reaux & Mazza, 2002), they nonetheless do not appear to appreciate the psychological aspect of 'seeing'. A series of experiments dedicated specifically to the concept 'seeing' investigated this issue. Prior to the actual experiments, the chimpanzees were trained to request food from an experimenter by making their species-typical begging gesture. During the actual experiments, the chimpanzees could request food from one of two experimenters. One of the experimenters could see the chimpanzees, whereas the other could not. Naturally,

if the chimpanzees understood 'seeing' they would gesture to the experimenter that could see them. For example, one experimenter had a blindfold over her mouth whereas the other had a blindfold over her eyes. In this case, the chimpanzee should gesture to the experimenter that has the blindfold over her mouth and not to the experimenter that has the blindfold over her eyes. However, only in one of three types of trial, where one experimenter was facing forward and the other had her back fully turned to the chimpanzees, did the chimpanzees spontaneously gesture to the right person. To test whether or not the chimpanzees were merely following a 'gesture to someone facing forward rule' during the experiment, a new trial-variation was designed. This time both experimenters had their back turned to the chimpanzees, but one of them was looking over her shoulder and could indeed see the chimpanzees. Surprisingly, the chimpanzees gestured at random to both experimenters. In time, the chimpanzees learned to perform the various trials at above chance levels. What exactly had these chimpanzees learned? Subsequent experiments led the researchers to conclude that the chimpanzees did not make inferences about 'seeing', but instead learned to use a prioritized set of observable features to determine if an experimenter was looking at them or not. Instead of relying on representations about the mental or attentional states of the experimenters, the chimpanzees appeared to follow three basic rules in the following fashion: (1) gesture to experimenters that are facing forward, (2) gesture to experimenters that have a visible face, and (3) gesture to experimenters that have their eyes open (Povinelli et al., 2000). For example, having a choice between one experimenter that is looking over her shoulder with her back turned and another that is facing the chimpanzees with her eyes closed, the chimpanzees will more frequently gesture to the latter experimenter. Based upon these experiments, one is led to think that reasoning about mental states is beyond the reach of these non-human great apes.

However, research by Hare et al. (2000; 2001) seems to show that chimpanzees do possess some understanding of 'seeing' (Hare et al., 2000) and even 'knowing' (Hare et al., 2001). They claim that begging for food is an unnatural situation for chimpanzees, since chimpanzees normally compete for food with their group mates and have a long evolutionary history of doing so (Hare & Wrangham, 2002). To solve this issue Hare et al. (2000) designed an experiment in which two chimpanzees, one dominant and one subordinate, would compete over food. For example, the subordinate could choose between two food sources: one food source was visible to both animals and the other was only visible to the subordinate. Their main finding was that the subordinate chimpanzees would much more often try to get the food that only they could see. Moreover, to make sure that the subordinates were not making their decision based on the behavior of the dominant, in a series of trials the subordinates had to make a choice before they could see any action on the side of the dominant. In another series of trials, the dominant's door was down to avoid the possibility that the subordinate's behavior was driven by intimidation or any other cues coming from the dominant. The same results were obtained as in previous sessions. The second series of studies obtained similar results (Hare et al., 2001) and provided evidence that chimpanzees can recall what a conspecific has and

has not seen in the immediate past. According to Hare et al. (2001) the subordinates' behavior was not based on any kind of learning, since it did not change during the course of experimentation. Overall, their data (Hare et al., 2000, 2001) suggests that, at least in competitive situations, the chimpanzees know what conspecifics do and do not see, and that they use this information strategically during social interactions. Subsequent studies by Hare and Tomasello (2004) have shown that chimpanzees are better at competitive tasks versus cooperative tasks. They claim that chimpanzee TOM skills are adaptations to competitive problem situations (Hare & Tomasello, 2004; Flombaum & Santos, 2005) and must therefore be understood in their socioecological context (Hare et al., 2001).

Similarly, Gómez (1998, 2004) has shown that a female orangutan is capable of understanding the mental state of 'knowing' versus 'not knowing' when a competitive element is added to the experimental set-up. The non-competitive/cooperative framework involved the following situation: Dona sits in her cage and in front of the cage there are two boxes locked with padlocks. The keys to the padlocks are kept in a different container. A 'baiter' comes in the room, takes the keys, opens one of the padlocks and places food in the box. A few seconds later the 'giver' comes in the room and asks Dona where the food is (or waits for her to make a request). When Dona points to one box, the 'giver' collects the keys, opens the padlock, gives the food to Dona and puts the keys back. This scenario is repeated several times. In the experimental trial, the 'baiter', after baiting the box, hides the keys in a hiding place in the room and leaves. If Dona understands the mental state of 'ignorance' or 'not knowing' she will point not only to the food, but also to the hiding place of the keys when the 'giver' comes into the room. The control condition involves hiding the keys in the presence of the 'giver', or the 'giver' himself hiding the keys, which would not evoke the same response of pointing both to the food and the hiding place. Dona failed six experimental trials and thus showed no sign of understanding at all. However, when a competitive element was introduced to the experimental set-up, Dona successfully passed the test. When the keys were hidden by a 'stranger' (unfamiliar to Dona) who entered the room after the 'baiter' had left, upon seeing the 'giver', she correctly pointed at both the food and the hiding place of the key in all seven experimental trials (Gómez, 1998, 2004).

Although these studies show some understanding of mental states in non-human great apes, neither Hare et al. (2001) nor Gómez (2004) claim that they possess human-like theory of mind skills. According to Hare et al. (2001), chimpanzees most likely possess what they refer to as "Level 1 perspective taking" and define as "knowing that others can see things that I cannot and vice versa", but might not have "Level 2 perspective taking" which they describe as "knowing exactly what others see, including that they see the same thing I do but from a different perspective" (Hare et al., 2001, p. 149). In their view, chimpanzees possess a representational understanding of others' behavior that allows them to foresee, remember and manipulate others' behavior. It might not allow them to recognize intentions in others, or understand how things might look from their perspective and so on. Although they most likely do not possess a full-blown human-like TOM system that

would enable them to 'place themselves in someone else's shoes', they do possess some TOM skills that humans and apes can be said to have in common. Indeed, recent work by Hare, Call and Tomasello (2006) suggests that chimpanzees possess, at least, an intermediate level of *intentional* deception characterized by actively concealing information from others. For example, a chimpanzee that is attempting to retrieve a prized food item that is placed in the vicinity of a human competitor, will actively conceal his behavior by initially moving out of the human competitor's sight, and hence, away from the food item, before proceeding to retrieve it. In a similar vein, Gómez (1996) argues that chimpanzees form representations about observable behaviors (what he refers to as 'covert' mental states). For example, when a gorilla infant sees a sweet being hidden under a cloth, the gorilla will remove the cloth and take the sweet. This shows that the gorilla 'knows' that the sweet is under the cover. Humans and non-human great apes may share the ability to form representations about observable behaviors (Gómez, 2004), but humans alone may be capable of forming representations about things that are not related to the observable world around us. For example, human four year-olds understand that another child may have a representation of the world that does not conform to the observable world around us, e.g. they understand false beliefs. Also, above and beyond actively concealing, human deception involves intentionally misleading (i.e. attempting to install a 'false belief' in a competitor).

Povinelli and Vonk (2003) make the distinction between "reasoning about behavior" on the one hand, and "reasoning about the mental states that underlie that behavior", i.e. human-like TOM, on the other. Based on the findings above, it is questionable if such a strict demarcation applies to non-human great apes. It nevertheless remains possible that, in line with Povinelli and Vonk's (2003) "behavioral abstraction hypothesis", chimpanzees' social skills are limited to reasoning about mere behavior. This entails that chimpanzees have the ability to (a) construct abstract categories of behavior, (b) make predictions about future behaviors that follow from past behaviors, and (c) adjust their own behavior accordingly. Though this is a plausible interpretation of the data, it is also plausible that chimpanzees share at least some TOM skills with humans, though they don't possess a full blown human-like TOM system (Tomasello, Call & Hare, 2003).

### 3. TOM and Introspection

If chimpanzees are able to reason about behavior, but not about mental states, or at least not to the extent that humans do, it appears that the difference lies in certain *additional* cognitive mechanism(s) that made human-like TOM skills possible. Decades ago Nicholas Humphrey (1986) claimed that the evolution of our 'inner eye' allowed for TOM. He argued that humans possess a kind of introspective awareness that allows them to understand others in much the same way as they understand themselves. Similarly, according to Gallup (1982) only those organisms that have the capacity to introspect can possess TOM skills such as empathy, intentional

deception, sorrow and the like. Moreover, Hauser (2000) claims, although he says further research might prove him wrong, that true empathy requires human-like self-awareness—the ability to understand one's own beliefs and feelings—which, according to him, non-human animals do not possess.

Can one claim that chimpanzees lack introspection? Although there is some evidence of self-awareness in chimpanzees (Gallup, 1970), there is no hard evidence for the existence of introspective abilities. Gallup (1982) maintains that the presence of mirror self-recognition (MSR) in chimpanzees is indicative of their introspective abilities. According to Gallup (1982), MSR reflects a sense of self that entails the existence of a 'mind'. 'Mind' in his definition is the monitoring of one's own mental states, in the sense of being able to distinguish between feelings of hunger, anger, fear, etc. In his view, evidence of 'mind' is evidence of introspection.

Gallup (1970) designed the 'mark test' to assess MSR in non-human animals. The 'mark test' consists of placing an odorless and tactile free mark on the eyebrow and above the opposite ear of anesthetized chimpanzees. Before the sedation, the chimpanzees had been exposed to a mirror for a prolonged time, allowing them to learn about their own reflection. After being fully recovered, the chimpanzees were observed for 30 minutes to account for spontaneous touching of the marked area (mirror absent condition). During the 30 minutes mirror absent condition after anesthesia, only one instance of mark-directed behavior was recorded. When the mirror was reintroduced, the chimpanzees' mark-directed behavior increased up to four or ten times (mirror condition). As a control, chimpanzees that were not mirror-exposed prior to being marked did not show mark-directed behaviors when exposed to a mirror (Gallup, 1970; Schilhab, 2004).

There are two parts to Gallup's theory: first, that the 'mark test' is a genuine indicator of MSR (Gallup, 1970); second, that MSR is indicative of introspection, and as such allows for introspectively based TOM skills such as empathy, deception and the like (Gallup, 1982). Gallup's MSR results have been reproduced numerous times (e.g., Lethmate & Dücker, 1973; Kitchen, Denton & Brent, 1996; Suarez & Gallup, 1981). Nevertheless, both claims remain contested in the literature (Heyes, 1994, 1998; Mitchell, 2002).

Several studies in the last decade have demonstrated persuasive evidence for the claim that the mark test gives insight into an animal's possession of MSR (Van Den Bos, 1999; Povinelli et al., 1997; Barth, Povinelli & Cant, 2004). The more controversial question remains whether MSR is an indicator of a more general self-concept and/or introspective abilities (Gallup, 1982). An alternative to Gallup's (1982) interpretation of the 'mark test' was brought forward by Robert Mitchell (2002): the "kinesthetic-visual matching" hypothesis (p. 345). Mitchell claims that the chimpanzees' behavior in front of the mirror and their apparent ability to recognize their own image, is based upon their "kinesthetic-visual matching" skills. Mitchell describes "kinesthetic-visual matching" as "the recognition of similarity between the feeling of one's own body's extent and movement (variously called 'kinesthesia,' 'somesthesia,' or 'proprioception') and how it looks (vision)" (Mitchell, 2002, p. 346). This ability to visualize one's own body in action or at least have a very general

idea of one's various body parts and their relative positions, allows chimpanzees to test the similarity between their own body and the mirror image, and consequently enables them to learn that the image in the mirror is like their own. According to Mitchell (2002) such kinesthetic-visual matching skills allow for MSR; they do not, however, entail the existence of a sense of self as outlined in Gallup's (1982) theory. Rather, they entail a kind of bodily awareness common to both chimpanzees and humans. In the same vein, Barth et al. (2004) claim that the 'mark test' is indicative of a kinesthetic representation of the self, but not of a self-concept or mental representation of the self *per se*. As is the case for human children, they defend the position that MSR reflects the ability to visually represent one's current body state and explicitly relate this image to the mirror image. Their research with human children revealed that MSR does not necessarily imply that a child has a self-concept that extends further than the present similarity between the child's own body image and the mirror image. Most children between 18 and 24 months show signs of MSR. They are able to use mirrors or live video feedback to locate and inspect marks on their face (Povinelli et al., 1996). However, children younger than four years of age do not reach above chance levels when performing a delayed self-recognition task. For example, in one of the delayed self-recognition experiments Povinelli et al. performed, the two- and four-year-old children are videotaped while an experimenter covertly places a sticker on their forehead during an unusual game. After the game, the children are shown the videotape of the game including the scene where the experimenter places a sticker on their forehead. Only the four year olds reach up to remove the sticker when the video revealed it being placed on their forehead. Although the younger children are generally capable of recognizing themselves in the delayed video, they do not seem to relate the delayed image to their present selves (Povinelli et al., 1996). This implies that MSR does not necessarily reflect a more general self-concept encompassing past, present and future. At the very least, there is no need to invoke a more general self-concept to explain the results obtained during the 'mark test' in both humans and chimpanzees. The recognition of the similarity between one's mirror image and one's kinesthetic body image is sufficient to explain the results.

Moreover, most children with autism succeed at the 'mark test' (Mitchell, 1997) although individuals with autism spectrum conditions (ASCs) typically have impaired self-awareness, i.e. impaired introspection (Hurlbert, Happé & Frith, 1994; Happé, 2003) and theory of mind impairments (Baron-Cohen, Happé & Frith, & Cohen, 1993; Frith & Happé, 1999). For example, children with autism typically fail false belief tasks (Baron-Cohen et al., 1993). Hurlbert et al. (1994) tested the introspective skills of three individuals with Asperger syndrome, showing their inner world to be very much unlike that of a normal adult. Reports normally obtained from participants included the following (singly or in combination): verbal inner experience, visual images, feelings (located in the body) and un-symbolized thinking (thoughts without words or pictures associated with them). In contrast, individuals with Asperger syndrome exclusively described visual inner experiences. In reference to Gallup's theory (1982) on MSR and TOM skills such as empathy and deception detection, it is important to mention that ASCs are associated with impaired

empathizing skills (Baron-Cohen, 1995; Baron-Cohen & Wheelwright, 2004), and that deception detection is deeply affected in these individuals (Grandin, 1995; Frith & Happé, 1999). For example, Temple Grandin (1995), an individual with autism, claims that unlike other individuals, she is ignorant to the subtle facial cues that may reveal deceitful intentions in others: "I had to learn to be suspicious, I had to learn it cognitively ... I couldn't see the *jealous* look on his face" (Grandin, 1995, p. 15). In conclusion, Gallup's (1982) claim that MSR is indicative of introspection and TOM is not supported (Mitchell, 1997).

#### 4. Is Introspection Relevant?

Assuming that introspection is uniquely human, the following question remains: how does introspection relate to TOM? There are two main theoretical views on how mindreading comes about, theory-theory (TT) and simulation theory (ST), along with several hybrid accounts involving both theory and simulation (see Carruthers & Smith, 1996; Goldman, 2006; Nichols & Stich, 2003). Rationality theory (RT) is a possible third option (Goldman, 2006). The main question for our purposes is, of course, if introspection plays a distinctive role in the attribution of mental states according to (any of) these views.

RT claims that individuals are rational agents acting according to rational principles. Basically, if an individual wishes to discern what someone thinks, wants, etc., he/she will infer this by appealing to what rationality dictates under such circumstances. Dennett (1987) defends an RT approach arguing that all mental state attribution, including self-attribution, is managed by rationality principles. Introspection is generally not part of the RT paradigm.

TT claims that the attribution of mental states occurs by virtue of "an implicitly held *theory* of the structure and functioning of the human mind" (Carruthers & Smith, 1996, p. 3). An example is the *child scientist* approach in which the child is seen as a scientist, modifying and reconstructing a theory of the world to best fit the facts at hand. In Gopnik's (1993) *child scientist* view, mental state attribution, including self-attribution, is achieved by means of theoretical inference. This is TT in its purest form. Introspection is not part of it. However, more diluted TT accounts might defend that third-person attribution relies on theoretical inference, while self-attribution is the result of self-monitoring (Goldman, 2006, p. 25). Generally, TT accounts do not attribute a role to introspection. Mental states are typically seen as "unobservable" or "abstract" and as such cannot be known or understood by means of introspection (Goldman, 2006, p. 259).

ST claims that mindreading does not originate from any kind of theory, but involves the ability to place ourselves in someone else's 'shoes' and imagine how the world looks like from their perspective. The observer simulates the mental word of the target by pretending to share his/her initial states, i.e. thoughts and emotions, and makes a decision based on those *pretend* states. These pretend states need not be arrived at consciously. They may be the result of emotional contagion, in which case

the concept *simulated* states is more applicable. After making a decision in the pretend mode, the observer predicts that this is the decision the target will make, and projects it onto the target. Goldman refers to this as "simulation-plus-projection" (Goldman, 2006, p. 40). 'Simulation' can be defined as the occurrence of a similar 'experiential' and/or psychological state in the observer as in the target, not acted upon by the observer. According to some ST accounts, introspection allows one to access these shared states (Carruthers & Smith, 1996).

Whereas Goldman's (1989, 2006) ST account defends the priority of introspection, Gordon (1996) defends a 'radical' simulationism that does not presuppose introspection. Goldman (2006) argues that "introspection is a common, indeed standard, way by which people recognize, discriminate, or detect the mental types of one's current mental tokens," (p. 259). As such, it allows us to distinguish between beliefs, desires, feelings, goals, aspirations, needs, etc.: "Third-person mindreading by simulation borrows classifications of one's own states to classify states of others" (p. 223). In line with traditional ST accounts, Goldman (2006) thus argues that first-person mindreading (i.e. mindreading one's own mind) and third-person mindreading (i.e. mindreading others' minds) both use introspection, introspection being *part* of the simulation-plus-projection routine involved in third-person mindreading. According to Gordon (1995, 1996), rather than a transfer or projection from self to other, simulation-based mindreading involves a *transformation* on behalf of the mindreader, much like actors who *become* the characters they play. As such, during mindreading, I do not consider what I would, think, do, etc. in your situation, but rather what *you*, the other, would think, do, etc. And, as Gordon (1995) puts it: "As the term 'introspection' is commonly understood, one can introspect only one's own mental states" (p. 57). Opposing Gordon's account, Goldman (2006) refers to empirical evidence linking introspection and third-person mindreading (e.g. Johnson et al., 2002; Kelley et al., 2002; Mitchell, Banaji & Macrae, 2005; Schmitz, Kawahara-Baccus & Johnson, 2004). These and other studies are discussed in §5.1.

Defending an introspection-based ST account, we similarly claim that, although simulation in itself need not involve mental state attribution, simulation-based *mindreading* nevertheless does. It requires an interpretation (implicit and/or explicit, un-verbalized and/or verbalized) of the shared psychological state of the observer and target in terms of mental state concepts, by means of introspection. Simulation itself reflects the occurrence of a shared experiential/psychological state between two individuals. It need not involve an interpretation of that shared state in terms of specific thoughts and feelings. Simulation-based mindreading does require such an interpretation. During simulation-based mindreading, the mindreader attributes specific thoughts and feelings to the target based upon his/her interpretation of this shared experiential/psychological state. As such, simulation-based mindreading requires the recognition of one's own mental states, and in humans, typically involves the explicit naming of those mental states during attribution.

A recent behavioral study on self-awareness and deception detection (Johnson et al., 2004) demonstrates that self-awareness and TOM are indeed

related abilities. The participants were presented with video-segments of various individuals either telling the truth or lying ('faking good' or 'faking bad') and were asked to indicate whether these individuals were truthful or deceitful. To test for possible self-awareness effects, the introspective skills of the participants were measured using the private self-consciousness scale (Fenigstein et al., 1975). The overall results indicated that the higher an individual's private self-consciousness, i.e. introspection, the better they are at detecting deception. Moreover, as discussed previously, studies on ASCs equally reveal a link between introspection and theory of mind (Hurlbert et al. 1994; Frith & Happé, 1999).

In the remainder of this paper, we will outline a cognitive neuroscience approach to mindreading congruent with introspection-based ST accounts. In overall defense of ST accounts, recent findings in cognitive neuroscience suggest that an individual's own experiences lie at the heart of their TOM skills (Gallese & Goldman, 1998; Goldman & Gallese, 2000). The premotor area F5 in macaque monkeys holds a newly discovered class of visuomotor neurons deemed mirror neurons (MNs) or F5 neurons: "Mirror neurons appear to form a cortical system matching observation and execution of goal-related motor actions" (Gallese & Goldman, 1998, p. 493). The same neurons fire when a monkey executes actions itself and when it observes similar actions in others. A comparable action/observing matching system has been found in humans comprising, at least, Broca's region (inferior frontal gyrus, IFG), the primary motor cortex, the superior temporal sulcus (STS) area and the parietal cortex (Iacoboni et al., 1999; Rizzolatti & Craighero, 2004). According to Gallese and Goldman (1998) MNs represent a primitive version or possibly a precursor in phylogeny of a simulation system that underlies human mindreading. MN activity seems to place an observer in the same 'mental shoes' as the target. Indeed, the creation of a similar state in the observer as witnessed in the target is in line with simulation approaches to mindreading that defend a kind of mental mimicry based on one's own experiences. As Goldman and Gallese (2000) mention, not every instance of mindreading necessarily requires a correspondent experience in oneself. Their view on simulation holds that mindreading skills ultimately rest on our own experiences, meaning that one can equally attribute mental states to others based on certain 'rules of thumb' that are derived from earlier cases of simulation (see Goldman & Gallese, 2000). Notwithstanding recent criticisms of ST accounts, this interpretation leaves room for non-mirroring simulation or a partially theory-based approach to mindreading, although not one involving the kind of *detached* theorizing or "cold methodology" (Gordon, 1996, p. 11) that is generally assumed in TT accounts.

Arguing against ST, Saxe (2005) claims that the errors that children and adults make when attributing mental states to others are not consistent with MN-accounts of mindreading, favoring a TT approach to mindreading instead. She mentions that four year olds do not understand the concept 'ignorance', that three year olds do not possess a complete theory regarding the sources of knowledge, and refers to the developmental time course of the errors children make (2005). For example, a three-year-old might not understand that individuals distinguish between a red and

a green ball by means of vision, but not by touching (source knowledge). Although young children typically possess beliefs and desires of their own around two years of age and typically correctly reason about others' desires at that time, it is not until at least one year later that they correctly reason about beliefs. According to Saxe, such a differential developmental time course for beliefs and desires must reflect a difference in the children's command of the *concepts* of belief and desire. According to TT, a child's mental state concepts are underpinned by their *theory* about the mind, and it is not until at least the third year of life that a child may acquire a full representational theory about the mind that captures the notion of misrepresentation, and hence allows the child to correctly reason about beliefs. However, according to Goldman and Sebanz (2005), this is not a convincing argument against ST, since the idea of two (or more) types of simulation (e.g. mirroring versus non-mirroring simulation) is in line with a differential time course for desire and belief understanding in the young child. More specifically, it opens up the possibility of one type of simulation developing earlier than the other. Strongly opposing Saxe's view are recent findings by Onishi and Baillargeon (2005) which show that children as young as 15 months of age show signs of implicit false belief understanding (as measured by their gaze behavior during a non-verbal 'Sally/Anne task'). There is substantial evidence that children's failure to pass explicit false belief tasks is due to limited processing (e.g. memory, inhibition) skills (for example, Mitchell & Lacohee, 1991; Saltmarsh et al., 1995); reflecting a processing rather than a conceptual change in the young child (Goldman, 2006).

According to Saxe (2005) adult's errors consist of systematically inaccurate and oversimplified beliefs about beliefs, involving mismatches between what we expect and what individuals actually do. We have overly optimistic expectations about the rationality of others' reasoning skills. Moreover, although we know that beliefs are sometimes false and that reasoning can be irrational, we attribute such lines of thinking more often to others than to ourselves. As a consequence, we tend to overestimate the prevalence of self-serving reasoning in others. Saxe (2005) claims that this congruence between people's beliefs about how the mind works and the actual mindreading errors they commit, gives us reason to believe these beliefs about others' minds are actually being used when we attribute mental states. She argues this is a problem for strong ST accounts, since simulators do not explicitly represent all minds functioning in the same way; rather, they simply derive information on how the mind works from their own mind.

This might indeed be a problem for ST accounts that rely solely on MN-based simulation processes. However, most, if not all current ST accounts of mindreading are not limited to MN-based processing. They may allow that *mirror-like* or *non-mirroring* simulation reflects the more cognitive aspects involved in mindreading (Focquaert & Platek, 2007; see 5.1.). Several fMRI studies have shown that areas outside of the human MNS are active during mindreading. It is therefore unlikely that ST in its strongest form, as Saxe (2005) argues against, provides an adequate account of human mindreading. ST accounts that do not exclusively rely on mirroring to explain mindreading, however, have no problem explaining errors in

mindreading. Furthermore, these accounts of adult simulation-based mindreading need only invoke an *attempt* to correctly replicate/attribute someone else's mental state by taking his/her perspective. Such perspective-taking undoubtedly suffers from systematic attribution errors. Considering that ST claims that this attempt originates from one's own standpoint, poor perspective-taking, reflecting the observer's 'egocentricity', is not unlikely. Last but not least, hybrid ST accounts allow for theorizing (non-simulation-based processing) to play a role in mindreading (Goldman & Sebanz, 2005).

## 5. Neurological Implications

### 5.1. Introspection-based Simulation Theory

Is there any reason to assume that self-awareness is related to TOM in the human brain, both at a MN level ('experiential' processing), and at more cognitive levels (introspection), of neuronal functioning? More specifically, do we find evidence of *mirroring* simulation, reflecting simulation-type processing in MN areas (i.e. motor and visceromotor regions containing MNs), as well as evidence of *mirror-like* simulation, reflecting simulation-type processing in non-MN brain areas (i.e. common activation in non-motor areas and/or common activation in non-motor single cells)? Undoubtedly, yes.

First of all, several imaging studies in normal individuals have revealed a mirroring mechanism that is implicated in social cognition and empathic processing, although the precise role of such a mirroring mechanism during explicit mindreading (i.e. mental state attribution) remains to be assessed (Goldman, 2006). According to Gallese, Keysers and Rizzolatti (2004), the mere observation of an action or emotion triggers the activation of the neural substrate, or at least part of it, involved when performing or experiencing that action or emotion oneself. Indeed, a study by Carr et al. (2003) showed robust activation in MN areas in normal individuals during the observation and imitation of emotionally expressive faces. (the *emotional face mirroring mechanism*). Moreover, Wicker et al. (2003) found common activation (anterior insula (AI) and anterior cingulate cortex (ACC)) for experiencing disgust oneself and observing disgust in others. (the *disgust mirroring mechanism*). Singer et al. (2004) conducted a study on pain-processing that revealed common activation in the ACC and AI when receiving a pain stimulus oneself and perceiving a similar pain stimulus in others (not directly observing, but anticipating). According to Gallese et al. (2004), Singer's study is in line with their visceromotor mirror matching theory of emotional understanding, since the AI and ACC are motor(-related) structures (the *pain mirroring mechanism*). Morrison, Lloyd and Roberts (2004) found similar results when comparing one's own pain experience to visual stimuli of someone else receiving a similar pain stimulus. Singer et al.'s (2004) study found that individuals' scores on two commonly used empathy measures was correlated with their ACC activation when perceiving someone else's pain. Moreover, the empathic processing was elicited in the absence of a direct emotional cue,

indicating the human ability to think about 'unobservables' (Povinelli & Vonk, 2006). A similar link between empathic processing and mirroring was found by Gazzola et al. (2006) when investigating the auditory mirror system in humans (the *auditory action mirroring mechanism*).

Second, several studies have shown that individuals with ASCs have introspection deficits and TOM impairments, while at the same time showing MN abnormalities both functionally and anatomically. Villalobos et al.'s (2005) findings suggest abnormalities in the frontal components of the dorsal stream in autistic individuals (abnormal functional connectivity between the primary visual cortex and inferior frontal lobe), which is consistent with the MN dysfunction hypothesis. Research on face processing in ASCs has shown abnormal processing in the STS compared to normal individuals (Pelphrey, Morris & McCarthy, 2005). The STS is indirectly connected to the ventral premotor cortex (vPM) by virtue of the inferior parietal lobe, which are all MN (or MN-related) areas. One of the major inputs to the vPM comes from the inferior parietal lobe, which is in turn reciprocally connected to the STS region (Gallese, 2001). In a recent fMRI study, children with ASCs did not show activation in the IFG while imitating or simply observing emotional faces. Moreover, their abnormality in MN functioning was inversely related to the level of social impairment manifested by each individual child (Dapretto et al., 2006). Oberman et al. (2005) found EEG evidence for MN impairments in ASCs: there was no suppression of presumed MN brain activity in the sensorimotor cortex while performing hand movements, while normal individuals show suppression both when performing hand movements and observing similar movements in others. In reference to anatomical abnormalities, Hadjikhani, Joseph, Snyder and Tager-Flusberg (2005) found significant thinning of MN areas in ASCs such as the IFG, the inferior parietal lobe and MNS-related areas such as the STS (the STS codes for movement observation, not execution; see Rizzolatti & Craighero, 2004). McAlonan et al. (2002) equally found volume decreases in frontal and parietal areas. A Diffusion Tensor Imaging (DTI) study looking at white matter structures in ASCs found reduced values in regions adjacent to the vPM cortex and the STS, which are MN or MNS-related regions, and also in the anterior cingulate gyri, the temporoparietal junctions, the temporal lobes approaching the amygdala bilaterally, the occipitotemporal tracts and the corpus callosum (Barnea-Goraly, 2004). If the social deficits found in ASCs can be related to MN impairments, this would imply a disruption or lack of 'experiential' processing (i.e. MN processing in premotor and parietal regions) that may lie at the root of their overall impairments in self-awareness and TOM. Although ASC brain abnormalities are not confined to MN areas (McAlonan et al., 2005), these areas do form a crucial component of their overall pathology, basically disrupting their ability for 'experiential' processing.

Third, in normal individuals several studies have shown that self-awareness and TOM comprise similar neurological activation patterns. An fMRI study by Ochsner et al. (2004) on emotional attribution showed that the attribution of emotions to self and other elicit shared neurological activation in several social brain areas. Participants in the study were asked either to identify the emotional response of

the central character depicted in photographs (pleasant, unpleasant or neutral), or to identify their own emotional response to each photograph, i.e. self condition, in comparison to control questions that concerned the location of the depicted scene (indoors, outside, or not sure). Common activation for self and other was found in the MPFC, lateral PFC, posterior cingulate cortex, and STS. Voegeley et al. (2001) found shared prefrontal activation for self and other mental state attribution. Participants were instructed to read four different kinds of short stories: physical stories (control condition); theory of mind stories (other condition); self stories involving theory of mind (self and other condition); self stories without theory of mind (self condition). Each story was followed by a question focusing on the specific nature of the story, e.g. physical, other-related, self- and other-related, and self-related. Common activation for the other condition and self-condition was found in the right prefrontal cortex. Gusnard, Akbudak, Shulman & Raichle (2001) and Johnson et al. (2002) found MPFC activation for introspective judgments and self-reflecting. Self-reflective thought activated the anterior regions of the MPFC. Reflecting upon current emotions also activates this region (Lane, Reiman, Ahern, Schwartz & Davidson, 1997; Gusnard et al., 2001). In a study by Kelley et al. (2002) participants were asked to make judgements about trait adjectives that were either self-relevant, other-relevant or case judgements. Both the self-condition and the other-condition elicited decreases in MPFC activity. This effect was more pronounced for other judgements resulting in stronger MPFC activity during the self-condition compared to the other-condition. Similarly, an fMRI study by Schmitz et al. (2004) revealed common MPFC activation for self-evaluation and other-evaluation (close friend or relative) of a stimulus set of thirty adjectives covering a broad range of personality traits (e.g. intelligent, shy). More recently, Mitchell et al. (2005) found activity in a region of the ventral MPFC that correlated with perceived self/other similarity, and more importantly, this effect was only seen for mental-state trials and not for non-mental-state trials. Frith and Frith (2003) mention that out of 12 mindreading studies that they reviewed, all showed MPFC activation. Posterior STS activation was found in 10 out of 12 mindreading studies. The MPFC has extensive connections to the STS, which is considered to be a MNS-related area. The MPFC region involved in these studies is the most anterior part of the paracingulate cortex which is often considered to be part of the anterior cingulate cortex (ACC) (Frith & Frith, 2003). Interestingly, the ACC proper is involved in attributing pain to self and others (Singer et al., 2004).

To summarize, we previously mentioned MN research in monkeys and humans that argues in favor of mirroring simulation-based mindreading. Although the precise role of mirroring simulation-type processing during mindreading remains to be assessed (Goldman, 2006), our first and second point further strengthens this claim by addressing mirroring mechanisms involved in normal social cognition and looking at MN deficits in ASCs. Our third point argues in favor of *mirror-like* simulation related to more cognitive neuronal activity in humans, highlighting the involvement of introspection or reflective thought during mindreading. Basically, we argue that, during mindreading, *mirroring* simulation contributes to a shared

experiential state between observer and target, whereas *mirror-like* simulation reflects cognitive mechanisms common to introspection and TOM. Although providing a strong case in favor of introspection-based simulation theory accounts of mindreading, the above-mentioned studies do not exclude the possibility of non-simulation-based mindreading.

### 5.2. Human and Chimpanzee Brains

Although not much is known about the developmental and evolutionary mechanisms underlying the emergence of human-specific brain features (Pollard et al., 2006), the recent publication of the initial sequence and analysis of the chimpanzee genome allows us to compare our own genome to that of our closest living relative (Sikela, 2006). Such knowledge may eventually allow us to identify the human-specific genomic changes underlying the evolution of the human brain and how these are related to human-specific brain anatomy and functioning. Genes that are potentially related to the evolution of uniquely human cognitive abilities, such as the FOXP2 gene involved in speech production (Enard et al., 2002) and the ASPM gene affecting brain size (Evans et al., 2004) are currently being identified (Pollard et al., 2006; Popesco et al., 2006; Sikela, 2006).

As mentioned, the human ACC is involved in emotional mindreading and adjacent to the MPFC (BA 32) which is implicated in mindreading. A special type of deep layer-neurons called spindle cell pyramidal neurons has been located in the ACC. These spindle cells are unique to humans and non-human great apes and are especially prominent in the human brain. They are not found in monkeys (Allman, Hakeem, Erwin, Nimchinsky & Hof, 2001, Hill & Walsh, 2005). The human spindle cells are more than twice as large as those in chimpanzees and bonobos, and three times as large as those in gorillas and orangutans. Histologically the ACC shows human-specific features, although the chimpanzee ACC gene expression profiles are closer to humans than to the gorilla or any other catarrhine primates (Uddin, Kaplan, Molnar-Szakacs, Zaidel & Iacoboni, 2005). Spindle cells have long-distance projections and are believed to be widely connected with diverse parts of the brain (Allman et al., 2001). The paracingulate cortex (BA 32), i.e. a region of the MPFC involved in mindreading tasks, although often included in the ACC, is deemed a cingulofrontal transition area sharing features of both cingulate and frontal cortices (Devinsky, Morrell & Vogt, 1995). Whether or not the recent evolutionary change in the ACC bears any relevance to the structure and function of this region in humans remains to be determined (Frith & Frith, 2003).

Taking into account that the human ACC and MPFC has been implicated in self-awareness and TOM, these brain regions might constitute part of the brain-network that is involved in human and non-human primate differences in mindreading. As mentioned, MN activity (IFG, STS, IPG) might constitute the basis of simulation-based mindreading, although it appears that more cognitive areas such as the MPFC and ACC, also operating by means of shared activation patterns

for self and other, i.e. simulation, are as much involved in this process as MN regions (Focquaert & Platek, 2007).

### 6. Conclusion

Now that the initial sequence of the chimpanzee genome is known, the focus of many researchers is on determining those genomic differences between chimpanzees and humans that set us apart. Establishing a link with phenotypic traits is still far from obvious, so it remains important to focus on both genotypic and phenotypic differences in the search for human-unique and chimpanzee-unique features. The presence of human-like TOM skills in non-human great apes has been a major focus of comparative cognitive research. Chimpanzees are, at the very least, capable of reasoning about behavior. Much less certainty exists about their ability to reason about mental states. Based upon the current findings, it is unlikely that chimpanzees possess a full-blown TOM system.

We propose that the difference in human and non-human mindreading lies in the human ability to introspect, which is presumably lacking in non-human great apes. Defending an introspection-based ST account of human mindreading, recent findings in cognitive neuroscience strongly suggest that the attribution of mental states to self and others is indeed intertwined anatomically and functionally, both at an 'experiential' level and a more 'cognitive' level of neuronal functioning. Future research should focus on the interplay between both levels and how this might contribute to human-specific mindreading.

### Acknowledgments

This research was funded by the Scientific Fund for Research, Flanders.

### References

- Allman, J.M., Hakeem, A., Erwin, J.M., Nimchinsky, E., & Hof, P. (2001). The anterior cingulate cortex. The evolution of an interface between emotion and cognition. *ANNALS of the New York Academy of Sciences*, 935, 107–117.
- Barne-Goraly, N., Kwon, H., Menon, V., Eliez, S., Lotspeich, L., & Reiss, A.L. (2004). White matter structure in autism: Preliminary evidence from diffusion tensor imaging. *Biological Psychiatry*, 55, 323–326.
- Baron-Cohen, S., Tager-Flusberg, H., & Cohen, D.J. (Eds.). (1993). *Understanding other minds: Perspectives from autism*. New York, NY: Oxford University Press.
- Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT Press.
- Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, 34, 164–175.
- Barth, J., Povinelli, D.J., & Cant, J.G.H. (2004). Bodily origins of SELF. In D. Beike, J.L. Lampinen, & D.A. Behrend (Eds.), *The self and memory* (pp. 11–43). New York, NY: Psychology Press.

- Byrne, R.W., & Whiten, A. (1990). Tactical deception in primates: the 1990 database. *Primate Report*, 27, 1-101.
- Carr, L., Iacoboni, M., Dubeau, M.C., Mazziotta, J.C., & Lenzi, G.L. (2003). Neural mechanisms of empathy in humans: A relay from neural systems for imitation to limbic areas. *Proceedings of the National Academy of Sciences*, 100, 5497-5502.
- Carruthers, P., & Smith, P.K. (Eds.). (1996). *Theories of theories of mind*. Cambridge, UK: Cambridge University Press.
- Dalton, K.M., Nacewicz, B.M., Johnstone, T., Schaefer, H.S., Gernsbacher, M.A., Goldsmith, H.H., et al. (2005). Gaze fixation and the neural circuitry of face processing in autism. *Nature neuroscience*, 8, 519-526.
- Dapretto, M., Davies, M.S., Pfeifer, J.H., Scott, A.A., Sigman, M., Bookheimer, S.Y., & Iacoboni, M. (2006). Understanding emotions in others: mirror neuron dysfunction in children with autism spectrum disorders. *Nature Neuroscience*, 9, 28-30.
- Dennett, D.C. (1987). *The intentional stance*. Cambridge, MA: MIT Press.
- Devinsky, O., Morrell, M.J., & Vogt, B.A. (1995). Contributions of the anterior cingulate cortex to behavior. *Brain*, 118, 279-306.
- de Waal, F.B.M., Thompson, E., & Proctor, J. (2005). Primates, monks and the mind. *Journal of Consciousness Studies*, 12, 38-54.
- Enard, W., Przeworski, M., Fisher, S.E., Lai, C.S.L., Wiebe, V., Kitano, T., et al. (2002). Molecular evolution of FOXP2, a gene involved in speech and language. *Nature*, 418, 869-872.
- Evans, P.D., Anderson, J.R., Vallender, E.J., Gilbert, S.L., Malcom, C.M., Dorus, S., & Lahn, B.T. (2004). Adaptive evolution of ASPM, a major determinant of cerebral cortical size in humans. *Human Molecular Genetics*, 13, 489-494.
- Fenigstein, A., Scheier, M. F., & Buss, A. H. (1975). Public and private self-consciousness: Assessment and theory. *Journal of Consulting and Clinical Psychology*, 43, 522-527.
- Flombaum, J.I., & Santos, L.R. (2005). Rhesus monkeys attribute perceptions to others. *Current Biology*, 15, 447-452.
- Focquaert, F., & Platek, S.M. (2007). Social cognition and the evolution of self-awareness. In S.M. Platek, J.P. Keenan, & T.K. Shackelford (Eds.), *Evolutionary cognitive neuroscience* (pp. 457-497). Cambridge, MA: MIT Press.
- Frith, U., & Happé, F. (1999). Theory of mind and self-consciousness: What is it like to be autistic? *Mind & Language*, 14, 1-22.
- Frith, U., & Frith, C.D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London B*, 358, 459-473.
- Gallese, V., & Goldman, A. (1998). Mirror neurons and the simulation theory of mind-reading. *TRENDS in Cognitive Sciences*, 2, 493-501.
- Gallese, V. (2001). The 'shared manifold' hypothesis. From mirror neurons to empathy. *Journal of Consciousness Studies*, 8, 33-50.
- Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *TRENDS in Cognitive Sciences*, 8, 396-403.
- Gallup Jr, G.G. (1970). Chimpanzees: self-recognition. *Science*, 167, 86-87.
- Gallup Jr, G.G. (1982). Self-awareness and the emergence of mind in primates. *American Journal of Primatology*, 2, 237-248.
- Gazzola, V., Aziz-Zadeh, L., & Keysers, C. (2006). Empathy and the somatotopic auditory mirror system in humans. *Current Biology*, 16, 1824-1829.
- Goldman, A.I. (1989). Interpretation psychologized. *Mind and Language*, 4, 161-185.
- Goldman, A.I., & Gallese, V. (2000). Reply to Schulkin. *TRENDS in Cognitive Sciences*, 4, 255-256.
- Goldman, A.I., & Sebanz, N. (2005). Simulation, mirroring, and a different argument from error. *TRENDS in Cognitive Sciences*, 9, 320.
- Goldman, A.I. (2006). Simulating minds. *The philosophy, psychology, and neuroscience of mindreading*. New York, NY: Oxford University Press.
- Gómez, J.C. (1996). Non-human primate theories of (non-human primate) minds: some issues concerning the origins of mind-reading. In P. Carruthers & P.K. Smith (Eds.), *Theories of theories of mind* (pp. 330-343). Cambridge, UK: Cambridge University Press.
- Gómez, J.C. (1998). Assessing theory of mind with nonverbal procedures: Problems with training methods and an alternative 'key' procedure. *Behavioral and Brain Sciences*, 21, 119-120.
- Gómez, J.C. (2004). *Apes, monkeys, children, and the growth of mind*. Cambridge, MA: Harvard University Press.
- Goodall, J. (1990). *Through a window: My thirty years with the chimpanzees of Gombe*. Boston, MA: Houghton Mifflin.
- Gopnik, A. (1993). How we know our minds: The illusion of first-person knowledge of intentionality. *Behavioral and Brain Sciences*, 16, 1-14.
- Gordon, R.M. (1995). Simulation without introspection or inference from me to you. In M. Davis & T. Stone (Eds.), *Mental simulation. Evaluations and applications* (pp. 53-67). Oxford, UK: Blackwell.
- Gordon, R.M. (1996). 'Radical' simulationism. In P. Carruthers & P.K. Smith (Eds.), *Theories of theories of mind* (pp. 11-21). Cambridge, UK: Cambridge University Press.
- Grandin, T. (1995). *Thinking in pictures and other reports from my life with autism*. New York, NY: Vintage Books.
- Gusnard, D.A., Akbudak, E., Shulman, G.L., & Raichle, M.E. (2001). Medial prefrontal cortex and self-referential mental activity: Relation to a default mode of brain function. *Proceedings of the National Academy of Sciences*, 98, 4259-4264.
- Hadjikhani, N., Joseph, R.M., Snyder, J., & Tager-Flusberg, H. (2006). Anatomical differences in the mirror neuron system and social cognition network in autism. *Cerebral Cortex*, 16, 1276-1282.
- Happé, F. (2003). Theory of mind and the self. *Annals of the New York Academy of Sciences*, 1001, 134-144.
- Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behavior*, 59, 771-785.
- Hare, B., Call, J., & Tomasello, M. (2001). Do chimpanzees know what conspecifics know? *Animal Behavior*, 61, 139-151.
- Hare, B., Call, J., & Tomasello, M. (2006). Chimpanzees deceive a human competitor by hiding. *Cognition*, 101, 495-514.
- Hare, B., & Wrangham, R. (2002). Integrating two evolutionary models for the study of social cognition. In M. Bekoff, C. Allen, & G.M. Burghardt (Eds.), *The cognitive animal* (pp. 363-369). Boston, MA: MIT Press.
- Hare, B., & Tomasello, M. (2004). Chimpanzees are more skilful in competitive than in cooperative cognitive tasks. *Animal Behavior*, 68, 571-581.
- Hauser, M.D. (2000). *Wild minds. What animals really think*. New York, NY: Henry Holt and Company.
- Heyes, C.M. (1994). Reflections on self-recognition in primates. *Animal Behavior*, 47, 909-919.
- Heyes, C.M. (1998). Theory of mind in nonhuman primates. *Behavioral and Brain Sciences*, 21, 101-148.
- Hill, R.S., & Walsh, C.A. (2005). Molecular insights into human brain evolution. *Nature*, 437, 64-66.
- Hurlbert, R.T., Happé, F., & Frith, U. (1994). Sampling the form of inner experience in three adults with Asperger's syndrome. *Psychological Medicine*, 24, 385-395.
- Humphrey, N. (1986). *The inner eye*. New York, NY: Oxford University Press.
- Iacoboni, M., Woods, R.P., Brass, M., Bekkering, H., Mazziotta, J.C., & Rizzolatti, G. (1999). Cortical mechanisms of human imitation. *Science*, 286, 2526-2528.
- Johnson, A.K., Barnacz, A., Constantino, P., Triano, J., Shackelford, T.K., & Keenan, J.P. (2004). Female deception detection as a function of commitment and self-awareness. *Personality and Individual Differences*, 37, 1417-1424.

- Johnson, J.C., Baxter, L.C., Wilder, L.S., Pipe, J.G., Heiserman, J.E., & Prigatano, G.P. (2002). Neural correlates of self-reflection. *Brain*, *125*, 1808–1814.
- Kelley, W.M., Macrae, C.N., Wyland, C.L., Caglar, S., Inati, S., & Heatherton, T.F. (2002). Finding the self: An event-related fMRI study. *Journal of Cognitive Neuroscience*, *14*, 785–794.
- Kitchen, A., Denton, D., & Brent, L. (1996). Self-recognition and abstraction abilities in the common chimpanzee studied with distorting mirrors. *Proceedings of the National Academy of Sciences USA*, *93*, 7405–7408.
- Lane, R.D., Reiman, E.M., Ahern, G.L., Schwartz, G.E., & Davidson, R.J. (1997). Neuroanatomical correlates of happiness, sadness, and disgust. *The American Journal of Psychiatry*, *154*, 926–933.
- Lethmate, J., & Dücker, G. (1973). Untersuchungen zum Selbsterkennen im Spiegel bei orang-utans und einigen anderen affenarten. *Zeitschrift für Tierpsychologie*, *33*, 248–269.
- Lewis, M. (2000). Self-conscious emotions: embarrassment, pride, shame and guilt. In M. Lewis & J.M. Haviland (Eds.), *Handbook of emotions* (2nd ed., pp. 623–636). New York, NY: Guilford Press.
- McAlonan, G.M., Daly, E., Kumari, V., Critchley, H.D., van Amelsvoort, T., Suckling, J., et al. (2002). Brain anatomy and sensorimotor gating in Asperger's syndrome. *Brain*, *127*, 1594–1606.
- McAlonan, G.M., Cheung, V., Cheung, C., Suckling, J., Lam, G.Y., Tai, K.S., et al. (2005). Mapping the brain in autism: A voxel-based MRI study of volumetric differences and intercorrelations in autism. *Brain*, *128*, 268–276.
- Mitchell, J.P., Banaji, M.R., & Macrae, C.N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience*, *17*(8), 1306–1315.
- Mitchell, P., & Lacoëe, H. (1991). Children's early understanding of false belief. *Cognition*, *39*, 107–127.
- Mitchell, R.W. (1997). A comparison of the self-awareness and kinesthetic-visual matching theories of self-recognition: Autistic children and others. *Annals of the New York Academy of Sciences*, *818*, 39–62.
- Mitchell, R.W. (2002). Kinesthetic-visual matching, imitation and self-recognition. In M. Bekoff, C. Allen, & G.M. Burghardt (Eds.), *The Cognitive Animal* (pp. 345–351). Boston, MA: MIT Press.
- Morrison, I., Lloyd, D., & Roberts, N. (2004). Vicarious responses to pain in anterior cingulate cortex: Is empathy a multisensory issue? *Cognitive, Affective & Behavioral Neuroscience*, *4*, 270–278.
- Nichols, S., & Stich, S. (2003). *Mindreading: An integrated account of pretense, self-awareness and understanding other minds*. Oxford, UK: Oxford University Press.
- Oberman, L.M., Hubbard, E.M., McCleery, J.P., Altschuler, E.L., Ramachandran, V.S., & Pineda, J.A. (2005). EEG evidence for mirror neuron dysfunction in autism spectrum disorders. *Cognitive Brain Research*, *24*, 190–198.
- Ochsner, K.N., Knierim, K., Ludlow, D.H., Hanelin, J., Ramachandran, T., Glover, G., & Mackey, S.C. (2004). Reflecting upon feelings: An fMRI study of neural systems supporting the attribution of emotion to self and other. *Journal of Cognitive Neuroscience*, *16*, 1746–1772.
- Onishi, K.H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs? *Science*, *308*, 255–258.
- Parr, L.A. (2001). Cognitive and physiological markers of emotional awareness in chimpanzees. *Pan troglodytes. Animal Cognition*, *4*, 223–229.
- Parr, L.A. (2003a). Emotional recognition by chimpanzees. In F. de Waal & P. Tyack (Eds.), *Animal social complexity: Intelligence, culture and individualized societies* (pp. 288–293). Cambridge, MA: Harvard University Press.
- Parr, L.A. (2003b). The discrimination of faces and their emotional content by chimpanzees (Pan troglodytes). *Annals of the New York Academy of Sciences*, *1000*, 56–78.
- Pelphrey, K.A., Morris, J.P., & McCarthy, G. (2005). Neural basis of eye gaze processing deficits in autism. *Brain*, *128*, 1038–1048.
- Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.A., Coppens, S., Pedersen, J.S., et al. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. *Nature*, *443*, 167–172.
- Popesco, M.C., Maclaren, E.J., Hopkins, J., Dumas, L., Cox, M., Meltesen, L., et al. (2006). Human lineage-specific amplification, selection, and neuronal expression of DUF1220 Domains. *Science*, *313*, 1304–1307.
- Povinelli, D.J., & Eddy, T.J. (1996). Chimpanzees: Joint visual attention. *Psychological Science*, *7*(3), 129–135.
- Povinelli, D.J., Landau, K.R., & Perilloux, H.K. (1996). Self-recognition in young children using delayed versus live feedback: Evidence of a developmental asynchrony. *Child Development*, *67*, 1540–1554.
- Povinelli, D.J., Gallup, G.G., Eddy, T.J., Bierschwale, D.T., Engstrom, M.C., Perilloux, H.K., & Toxopeus, I.B. (1997). Chimpanzees recognize themselves in mirrors. *Animal Behavior*, *53*, 1083–1088.
- Povinelli, D.J., Bierschwale, D.T., & Cech, C.G. (1999). Comprehension of seeing as a referential act in young children, but not juvenile chimpanzees. *British Journal of Developmental Psychology*, *17*, 37–60.
- Povinelli, D.J., Bering, J.M., & Giambrone, S. (2000). Toward a science of other minds: Escaping the argument by analogy. *Cognitive Science*, *24*, 509–541.
- Povinelli, D.J., Dunphy-Lelli, S., Reaux, J.E., & Mazza, M.P. (2002). Psychological diversity in chimpanzees and chimpanzees: New longitudinal assessments of chimpanzees' understanding of attention. *Brain, Behavior & Evolution*, *59*, 33–53.
- Povinelli, D.J., & Vonk, J. (2003). Chimpanzee minds: Suspiciously human? *TRENDS in Cognitive Sciences*, *7*, 157–160.
- Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*, 515–526.
- Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, *27*, 169–192.
- Saltmarsh, R., Mitchell, P., & Robinson, E. (1995). Realism and children's early grasp of mental representation: Belief-based judgements in the state change task. *Cognition*, *57*, 297–325.
- Santos, L.R., Flombaum, J.I., & Phillips, W. (2007). The evolution of human mindreading: How nonhuman primates can inform social cognitive neuroscience. In S.M. Platek, J.P. Keenan, & T.K. Shackelford (Eds.), *Evolutionary cognitive neuroscience* (pp. 433–456). Cambridge, MA: MIT Press.
- Saxe, R. (2005). Against simulation: The argument from error. *TRENDS in Cognitive Sciences*, *9*, 174–179.
- Schilhab, T.S.S. (2004). What mirror self-recognition in nonhumans can tell us about aspects of the self. *Biology and Philosophy*, *19*, 111–126.
- Schmitz, T.W., Kawahara-Baccus, T.N., & Johnson, S.C. (2004). Metacognitive evaluation, self-relevance, and the right prefrontal cortex. *NeuroImage*, *22*, 941–947.
- Sikela, J.M. (2006). The jewels of our genome: The search for the genomic changes underlying the evolutionarily unique capacities of the human brain. *PLoS Genetics*, *2*, 646–655.
- Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R.J., & Frith, C.D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science*, *303*, 1157–1162.
- Suarez, S.D., & Gallup Jr, G.G. (1981). Self-recognition in chimpanzees and orangutans, but not gorillas. *Journal of Human Evolution*, *10*, 175–188.
- Tomasello, M., Call, J., & Hare, B. (2003). Chimpanzees understand psychological states—the question is which ones and to what extent. *TRENDS in Cognitive Sciences*, *7*, 153–156.

- Uddin, L.Q., Kaplan, J.T., Molnar-Szakacs, I., Zaidel, E., & Iacoboni, M. (2005). Self-face recognition activates a frontoparietal 'mirror' network in the right hemisphere: An event-related fMRI study. *NeuroImage*, *25*, 926-935.
- Van Den Bos, R. (1999). Reflection on self-recognition in nonhuman primates. *Animal Behavior*, *58*, F1-F9.
- Villalobos, M.E., Mizuno, A., Dahl, B.C., Kemmotsu, N., & Müller, R.-A. (2005). Reduced functional connectivity between V1 and inferior frontal cortex associated with visuomotor performance in autism. *NeuroImage*, *25*, 916-925.
- Vogele, K., Bussfeld, P., Newen, A., Herrmann, S., Happé, F., Falkai, P., et al. (2001). Mind reading: Neural mechanisms of theory of mind and self-perspective. *NeuroImage*, *14*, 170-181.
- Vonk, J., & Povinelli, D.J. (2006). Similarity and difference in the conceptual systems of primates: The unobservability hypothesis. In E. Wasserman & T. Zentall (Eds.), *Comparative cognition: Experimental explorations of animal intelligence* (pp. 363-387). Oxford, UK: Oxford University Press.

## What is Self-Control?

Edmund Henden

*What is self-control and how does the concept of self-control relate to the notion of will-power? A widespread philosophical opinion has been that the notion of will-power does not add anything beyond what can be said using other motivational notions, such as strength of desire and intention. One exception is Richard Holton who, inspired by recent research in social psychology, has argued that will-power is a separate faculty needed for persisting in one's resolutions, what he calls 'strength of will'. However, he distinguishes strength of will from self-control. In this paper I argue that will-power is essential also to a certain form of self-control. I support this claim by arguments showing that the traditional philosophical accounts of self-control run into difficulties because they pay insufficient attention to will-power as an independent source of motivation.*

**Keywords:** *Acrasia; Intention; Motivation; Self-Control; Will-Power*

Will-Power is to the mind like a strong blind man who carries on his shoulders a lame man who can see.

(Arthur Schopenhauer)

A central concept in the debate about responsible agency is the concept of self-control. But what is self-control and how does it work? Roughly, we can distinguish three different approaches to self-control in the philosophical literature, what I shall call 'the desire account', 'the cognitive-dispositional account' and 'the volitional account'. While the first two accounts explain self-control in terms of either a special kind of desire or style of thinking, the third explains it in terms of a volition or act of will. Curiously, one notion that appears to have been absent in the ensuing debate between these accounts is the notion of *will-power*. I use the word 'curiously' here because from a commonsense point of view, will-power is exactly

---

Edmund Henden is a post-doctoral fellow in the Department of Philosophy, Classics, History of Art and Ideas, University of Oslo, where he is affiliated with the Centre for the Study of Mind in Nature (CSMN). Correspondence to: Edmund Henden, Dept. of Philosophy, Classics, History of Art, and Ideas, 1020 Blindern, Oslo 0315, Norway. Email: edmund.henden@ifikk.uio.no