Decoupling beliefs from reality in the brain: an ERP study of theory of mind

David Liu,[1,CA] Mark A. Sabbagh,[3] William J. Gehring[1] and Henry M. Wellman[1,2]

[1]Department of Psychology and [2]Center for Human Growth and Development, University of Michigan, 525 E University, Ann Arbor, MI 48109-1109, USA; [3]Department of Psychology, Queen's University, Kingston, Ontario, Canada

[CA]Corresponding Author: davidliu@umich.edu

Received 13 January 2004; accepted 23 January 2004

DOI: 10.1097/01.wnr.0000123388.87650.06

Theory of mind, attributing behaviors to mental states, is a cognitive ability central to human social interactions. To investigate the neural substrates of theory of mind reasoning, we recorded human event-related brain potentials (ERP) while participants made judgments about belief and judgments about reality. A late ERP component (peaking around 800 ms post-stimulus) with a left frontal scalp distribution, which was inconsistent with a source in the anterior paracingulate cortex and consistent with a source possibly in the left orbitofrontal cortex, differentiated judgments about belief and about reality. This late left frontal component is probably associated with the decoupling mechanism that distinguishes mental states from reality. *NeuroReport* 15:991–995 © 2004 Lippincott Williams & Wilkins.

Key words: Event-related potentials; Prefrontal cortex; Theory of mind

INTRODUCTION

Social cognition of some sort characterizes many species. Many scholars believe that what sets human social cognition uniquely apart from that of other animals, even other primates, is having a theory of mind; that is, construing behaviors as caused by mental states such as beliefs, desires, intentions, and emotions [1,2]. A hallmark of this mentalizing ability is evident in reasoning about false beliefs. When children, and adults, predict mistaken behavior on the basis of a person's false belief, there is evidence that they understand actions as based on actors' representation of the world rather than on the reality of the world itself. In typically developing children, this ability emerges between the ages of 3 and 6 years [3]. However, theory of mind is severely and specifically impaired in individuals with autism, and this has led researchers to suggest that there may be distinct neural systems that support reasoning specifically about mental states [4,5].

In recent years, functional neuroimaging has been used to identify the brain structures associated with this distinctively human ability. In a recent review of PET and fMRI studies, Gallagher and Frith [6] argued that the superior temporal sulcus and the temporal poles are associated more generally with social cognition but it is the medial prefrontal cortex, particularly the anterior paracingulate cortex, that is the key region for theory of mind. The critical role of the prefrontal cortex is further supported by neuropsychological studies that show frontal lobe damage associated with impairments in theory of mind [7–11], including the medial prefrontal cortex as well as the orbitofrontal cortex. The orbitofrontal cortex has been hypothesized by some to be a vital component of the neural circuit for theory of mind [12,13].

In most conceptual analyses of theory of mind, in order to reason about mental states, one must separate or decouple mental states from reality (e.g. appreciate simultaneously that someone else believes that the apple is in the cupboard although it is really in the refrigerator). Therefore, when Gallagher and Frith [6] concluded that functional neuroimaging studies point to the anterior paracingulate cortex as critically involved in theory of mind, they argued that it was the neural source of the decoupling processes that distinguish mind and world. Surprisingly, however, no previous neuroimaging or electrophysiological studies of theory of mind have actually tested this decoupling contrast between judgments about belief and judgments about reality. The current study investigated this question directly. In particular, we recorded human event-related brain potentials (ERP) while participants made judgments about belief and judgments about reality.

While functional neuroimaging studies are beginning to localize regions of neural circuitry implicated in theory of mind, as a whole they have several limitations. These studies typically index reflective rather than real time, or on-line, mentalizing processes, and they provide little information about the temporal nature of neural processes associated with theory of mind. Most neuroimaging studies of theory of mind have used off-line tasks, which asked participants to reflect on different scenarios and then, outside of the scanner, retrospectively explain people's behaviors in the scenarios. Only two neuroimaging studies have acquired scans based on real time mentalizing judgments [14,15]. Therefore, studies on the neural basis of mentalizing should target on-line processes. A related matter is that neuroimaging studies using PET and fMRI, even the two that used on-line judgment tasks, are restricted by the poor temporal resolution of these techniques. However, the neural mechanisms underlying theory of mind are hypothesized to involve fast, on-line decoupling processes, necessary for people to successfully navigate the quick and complex nuances of social interactions. Under some proposals, inferring mental states should be very fast indeed, and automatic [16]. Therefore, tracking the temporal nature of theory of mind neural processing is needed to validate the neuroimaging results and to better understand the neural mechanisms involved. Studies using human ERP, which have very precise temporal resolutions, are well suited to this task.

To date, there has been only one ERP study designed to index the neural systems associated with reasoning about mental states. Sabbagh and Taylor [17] conducted an ERP study that asked participants to reason about false-belief representations *vs* false-photograph representations. In both conditions, participants had to make judgments about out of date and thus false representations: the only difference between conditions was whether the representations were mental (beliefs) or non-mental (photographs). Results showed that reasoning about these two types of representations were differentiated by a slow, late component of the ERP (peaking around 800 ms post-stimulus) with left frontal scalp distribution. Sabbagh and Taylor's [17] study merits replication and extension. Most important, the current study examined whether the late left frontal component that differentiated reasoning about false-belief representations and false-photograph representations also differentiates reasoning about beliefs and reality. Establishing this would strengthen the case that the late left frontal component is an index of theory of mind reasoning, *per se*. Second, the medial prefrontal cortex findings from functional neuroimaging studies are not generally lateralized. The observation of left frontal scalp distribution from Sabbagh and Taylor's ERP study may be related to different theory of mind processes than those that have been tapped in neuroimaging studies or it may be related to the text stimuli used in their study. It has been suggested that text-based theory of mind tasks are more likely to reveal left-lateralized effects [18]. In the current study, we presented animated vignettes, verbally narrated, as stimuli, rather than text. Lastly, Sabbagh and Taylor did not perform source localization analysis of their data and thus, were unable to make strong hypotheses about the cortical regions responsible for their effects. A particular focus of the current study was determining whether reasoning about beliefs *vs* reality is associated with ERP effects attributable to the medial prefrontal cortex.

MATERIALS AND METHODS

Subjects: Seventeen adult volunteers (age 19–35 years; five males and 12 females) participated in the study. One female was left-handed. All participants had normal or corrected-to-normal vision. The Institutional Review Board of the University of Michigan approved the study, and all participants gave written informed consent.

Stimuli: Forty trials of similar cartoon animations were presented on a computer monitor to participants. The structure of all 40 trials was the same. In each trial (Fig. 1), a cartoon character (e.g. Garfield) begins by standing next to two boxes and puts an animal into each box (two animals total). The cartoon character then walks in front of the boxes so that he or she cannot see either box. Then one of the animals in the boxes jumps out of the box and either moves to the other box (30 trials) or goes back into the same box (10

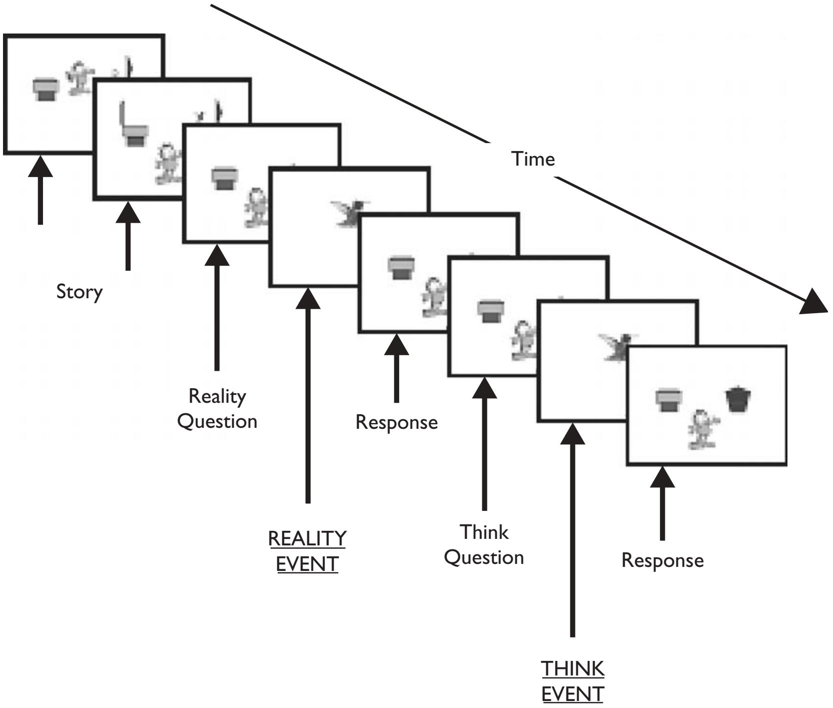

Fig. 1. Timeline of stimuli. After the story portion of each trial, the participants made a reality judgment and a think judgment. Reality event and think event were the target events to which the reality condition and the think condition ERP data were time-locked, respectively.

trials). After the story portion of each trial, participants were then asked by the experimenter to make a reality judgment and a think judgment. The order of the reality and think judgments was counter-balanced between subjects. For a reality judgment, participants were asked to judge where one of the two animals was in reality ('Really, where is this?'), followed by the presentation of one of the two animals. This pictorial presentation of a single animal (labeled in Fig. 1 as reality event) was the target event to which the reality condition ERP data were time-locked. For a Think judgment, participants were asked to judge where the cartoon character thinks one of the two animals is (e.g. 'Where does Garfield think this is?'), followed by the presentation of one of the two animals. This pictorial presentation of a single animal (labeled in Fig. 1 as think event) was the target event to which the think condition ERP data were time-locked.

Procedure: After participants filled out a consent form, the electrodes and electrode cap were applied, and the participants were seated in the recording booth. For each trial, participants were presented with the story portion and then asked, verbally, by the experimenter to make a reality judgment and a think judgment. Each ERP eliciting stimulus was presented for 2000 ms. After the offset of the ERP eliciting stimulus, participants provided their answers verbally, and the experimenter recorded their responses on the computer.

ERP recording and analysis: The EEG was recorded from tin electrodes embedded in a nylon mesh cap (Quik-cap, Neuromedical Supplies, Sterling, Virginia, USA). EEG data from 40 channels were recorded with a left mastoid reference and a forehead ground. An average mastoid reference was derived off-line using right mastoid data. The electrooculogram (EOG) was recorded from tin electrodes above and below the right eye and external to the outer canthus of each eye. Impedance was kept below 10 kΩ. EEG and EOG were amplified by SYNAMPS DC amplifiers (Neuroscan Labs, Sterling, Virginia, USA) and filtered on-line from 0.01 Hz to 70 Hz (half-amplitude cutoff). The data were digitized at 250 Hz. EEG data were corrected for ocular movement artifacts using the algorithm of Gratton *et al.* [19]. Prior to analysis, the data were filtered with a 9-point Chebyshev type II low-pass digital filter, with a half-amplitude cutoff at 12 Hz.

RESULTS

All participants completed all 40 trials, and performance on reality judgments and performance on think judgments were equally high (100% correct for both conditions). Figure 2 shows the ERP waveforms from the grand average of all subjects for reality and think judgments from nine electrodes in a 3 × 3 grid encompassing electrodes from front to back (caudality) and from left to right (laterality) on the scalp. Visual inspection of the waveforms suggests that a late divergence between reality and think waveforms peaked, as in Sabbagh and Taylor's [19] ERP study, around 800 ms post-stimulus.

The peak for each type of judgment's late ERP component was quantified as the mean amplitude in the 700–900 ms epoch following the onset of the target stimulus, relative to a 100 ms pre-stimulus baseline. A 2 (condition: reality *vs* think) × 3 (caudality) × 3 (laterality) repeated measures ANOVA was conducted on the 3 × 3 grid of electrodes in

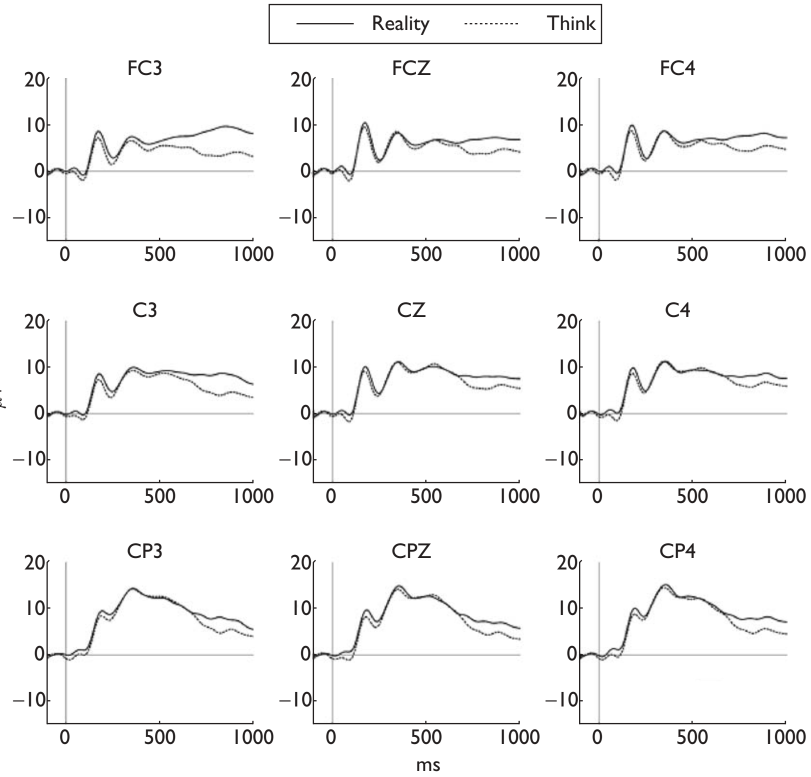

Fig. 2. Grand average ERP waveforms for reality (solid lines) and think (dotted lines) conditions from nine electrodes in a 3 × 3 grid encompassing electrodes from front to back (top to bottom) and from left to right on the scalp.

Fig. 2. In this analysis, when necessary, *p*-values were adjusted using the Greenhouse-Geisser correction for violations of the ANOVA assumption of sphericity. There were no significant main effects for condition, caudality, and laterality. There was a two-way interaction effect between condition and caudality (F(2,32)=4.13, $p < 0.05$, MSE=3.71), and there was a two-way interaction effect between condition and laterality (F(2,32)=5.11, $p < 0.05$, MSE=3.42). Most important, there was a three-way interaction between condition, caudality, and laterality (F(4,64)=4.66, $p < 0.005$, MSE=1.04). These interaction effects show that the late divergence between reality and think waveforms had a left frontal scalp distribution. That is, the difference between reality and think electrophysiological potentials were greatest in left frontal electrodes while the potentials did not differ greatly in posterior or right hemisphere electrodes. This is further illustrated by the topographic map of scalp electrical activity, shown in Fig. 3, which is computed as the mean amplitude difference between the conditions (reality subtracted from think) in the 700–900 ms epoch. The results suggest that reasoning about mental states *vs* reality is associated with a late ERP component in left frontal electrodes.

It appears that the late left frontal component, which was associated with the contrast between on-line judgments about belief and on-line judgments about reality, was generated from a single source within the left frontal cortex. This left frontal scalp distribution of the ERP activity that distinguishes the reality and think conditions is similar to the distribution found in Sabbagh and Taylor's [19] ERP study. To gain further information about the source of this late left frontal component, we used the brain electrical source analysis (BESA) software [20] to derive a best-fit, single-dipole model based on the mean amplitude difference between the conditions in the 700–900 ms epoch of all 40 channels. Figure 4 displays the best-fit dipole solution. The residual (unaccounted for) variance associated with the best-fit dipole model was less than 13%. The dipole modeling identified a source inconsistent with anterior paracingulate cortex and more consistent with left orbitofrontal cortex (x=−21.56, y=26.79, z=−14.76). It appears that the generator of the late left frontal component is ventral and left lateral of the anterior paracingulate cortex.

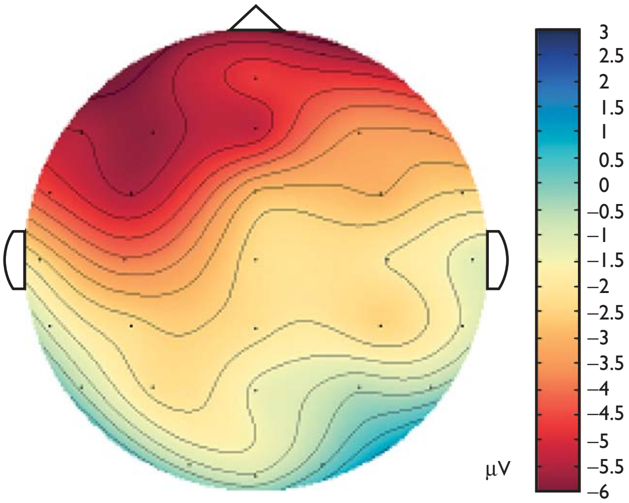

Fig. 3. Topographic voltage map of scalp electrical activity of the mean amplitude difference between the conditions (reality subtracted from think) in the 700–900 ms epoch.

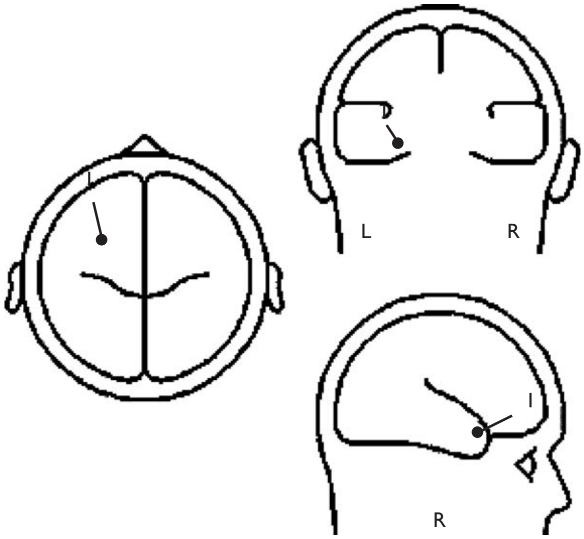

Fig. 4. BESA software [20] derived best-fit, single-dipole solution based on the mean amplitude difference between the conditions in the 700–900 ms epoch. The dipole modeling identified a source inconsistent with anterior paracingulate cortex and more consistent with left orbitofrontal cortex (x=−21.56, y=26.79, z=−14.76).

DISCUSSION

The current study is the first study directly targeting the cortical activity associated with reasoning about mental states *vs* reality. The results demonstrated that waveforms diverged between judgments about belief and judgments about reality in left frontal electrodes and peaked around 800 ms post-stimulus, agreeing with and extending the findings of Sabbagh and Taylor's [17] ERP study. The temporal characteristics and the scalp distribution of the late left frontal component observed in the current study are very similar to their findings. There is a difference in polarity between the studies: the belief waveform more positive than the photograph waveform in Sabbagh and Taylor's study, but the reality waveform more positive than the belief waveform in the current study. However, given that the contrasts in the two studies were different, it is not necessarily the case that the polarity of the ERP effect should be the same in the two studies. The computations that distinguish the belief condition from the reality condition in our study may differ from those that distinguish the belief condition from the photograph condition in Sabbagh and Taylor's study. Thus, taken together, the two studies provide support for some aspect of theory of mind being indexed by the left frontal ERP component, and additional studies exploring these conditions in more detail will be necessary to determine how specific computations involved in theory of mind reasoning relate to the polarity of the scalp effect.

In contrast to Sabbagh and Taylor's use of text, we presented cartoon animation stimuli to minimize the possibility of theory of mind processing interacting with text processing. Nevertheless, we too found left frontal scalp distribution. This left frontal scalp localization of the late ERP component was confirmed with source localization using BESA and was consistent with a source in the left orbitofrontal cortex. Because the late left frontal component was generated in a brain region distinct from left frontal regions primarily associated with linguistic processing (i.e. dorsolateral prefrontal cortex and Broca's area), it appears unlikely that the ERP effects observed were due to an interaction with text processing. Both the time course and the spatial distribution of our findings are noteworthy. On-line theory of mind processing has been proposed to be fast and automatic; some researchers have even argued that mentalizing is modularized and perception-like in its speed and automaticity [16]. The late ERP component that differentiated judgments about belief and about reality was fast and on-line, occurring in < 900 ms. However, with a late peak and extended time course, it does not seem likely that the late left frontal component is perception-like in its automaticity.

As mentioned in the introduction, Gallagher and Frith [6] argued that the anterior paracingulate cortex is critically involved in the decoupling mechanism that distinguishes mental states from reality. However, according to the BESA solution for our results, the source of the late left frontal component is in a different brain region than the anterior paracingulate cortex. These findings seem particularly inconsistent with Gallagher and Frith's proposal because our experimental task specifically contrasts on-line judgments about belief and about reality. One possible explanation for the inconsistency is that the functional neuroimaging studies that have found anterior paracingulate cortex activation did not actually examine the fast, on-line decoupling mechanism because they did not use tasks that precisely contrasted judgments about mental states and judgments about reality and because of the limited temporal resolutions of PET and fMRI.

Gallagher and Frith [6] argued that the orbitofrontal cortex, along with the superior temporal sulcus, the temporal poles, and the amygdala, is part of a neural network of the social brain circuit [21] responsible for social cognition in general but not for theory of mind in particular. Patients with damage to their orbitofrontal cortex and ventromedial cortex have been found to behave in socially inappropriate manners and to be poor at affective decision-making [22]. However, Baron-Cohen and Ring [12] and Brothers and Ring [13] argued that the orbitofrontal cortex is critical for theory of mind. A single positron emission computerized tomography (SPECT) study of theory of mind has found activation of the orbitofrontal cortex (right lateralized) associated with recognition of mental state terms [23]. Furthermore, two neuropsychological studies have found bilateral orbitofrontal cortex damage associated with impairment in theory of mind [8,10]. However, the role of left *vs* right orbitofrontal cortex requires additional research. These sorts of findings, bolstered by our current study, suggest that the orbitofrontal cortex is possibly critical for theory-of-mind reasoning. Further, they raise the question whether the late frontal ERP activity observed in the current study is part of the same mentalizing neural circuitry as the activity observed in the anterior paracingulate cortex in functional neuroimaging studies.

CONCLUSION

A late ERP component with a left frontal scalp distribution, which was consistent with a source possibly in the left orbitofrontal cortex, differentiated judgments about belief and about reality. We conclude that this late left frontal component is probably associated with the decoupling mechanism that distinguishes mental states from reality. Further electrophysiological and functional neuroimaging studies using comparable tasks are needed to examine the relation between the orbitofrontal cortex and the anterior paracingulate cortex in theory of mind.

REFERENCES

1. Tomasello M. *The Cultural Origins of Human Cognition*. Cambridge, MA: Harvard University Press, 1999.
2. Wellman HM. *The Child's Theory of Mind*. Cambridge, MA: MIT Press, 1990.
3. Wellman HM, Cross D and Watson J. Meta-analysis of theory-of-mind development: the truth about false belief. *Child Dev* **72**, 655–684 (2001).
4. Baron-Cohen S. *Mindblindness: an Essay on Autism and Theory of Mind*. Cambridge, MA: MIT Press, 1995.
5. Frith CD and Frith U. Interacting minds – a biological basis. *Science* **286**, 1692–1695 (1999).
6. Gallagher HL and Frith CD. Functional imaging of theory of mind. *Trends Cogn Sci* **7**, 77–83 (2003).
7. Channon S and Crawford S. The effects of anterior lesions on performance on a story comprehension test: left anterior impairment on a theory of mind-type task. *Neuropsychologia* **38**, 1006–1017 (2000).
8. Happe F, Malhi GS and Checkley S. 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 (2001).
9. Rowe AD, Bullock PR, Polkey CE and Morris RG. 'Theory of mind' impairments and their relationship to executive functioning following frontal lobe excisions. *Brain* **124**, 600–616 (2001).
10. Stone VE, Baron-Cohen S and Knight RT. Frontal lobe contributions to theory of mind. *J Cogn Neurosci* **10**, 640–656 (1998).
11. Stuss DT, Gallup GG and Alexander MP. The frontal lobes are necessary for theory of mind. *Brain* **124**, 279–286 (2001).
12. Baron-Cohen S and Ring H. A model of the mindreading system: neuropsychological and neurobiological perspectives. In: Mitchell P and Lewis C (eds). *Origins of an Understanding of Mind*. Hillsdale, NJ: Erlbaum; 1994, pp. 183–207.
13. Brothers L and Ring B. A neuroethological framework for the representation of minds. *J Cogn Neurosci* **4**, 107–118 (1992).
14. Gallagher HL, Jack AI, Roepstorff A and Frith CD. Imaging the intentional stance in a competitive game. *Neuroimage* **16**, 814–821 (2002).
15. McCabe K, Houser D, Ryan L, Smith V and Trouard T. A functional imaging study of cooperation in two-person reciprocal exchange. *Proc Natl Acad Sci USA* **98**, 11832–11835 (2001).
16. Leslie AM. Pretending and believing: issues in the theory of ToMM. *Cognition* **50**, 211–238 (1994).
17. Sabbagh MA and Taylor M. Neural correlates of theory-of-mind reasoning: an event-related potential study. *Psychol Sci* **11**, 46–50 (2000).
18. Gallagher HL, Happe F, Brunswick N, Fletcher PC, Frith U and Frith CD. Reading the mind in cartoons and stories: an fMRI study of 'theory of the mind' in verbal and nonverbal tasks. *Neuropsychologia* **38**, 11–21 (2000).
19. Gratton G, Coles M and Donchin E. A new method for off-line removal of ocular artifact. *Electroencephalogr Clin Neurophysiol* **55**, 468–484 (1983).
20. Scherg M. Fundamental of dipole source potential analysis. In: Grandori F, Hoke M, Romani GL (eds). *Advances in Audiology: Vol. 6. Auditory Evoked Magnetic Fields and Electrical Potentials*. Basel, Switzerland: Karger; 1990, pp. 40–69.
21. Brothers L. The social brain: a project for integrating primate behaviour and neurophysiology in a new domain. *Concepts Neurosci* **1**, 27–51 (1990).
22. Bechara A, Damasio H and Damasio AR. Emotion, decision making and the orbitofrontal cortex. *Cerebr Cortex* **10**, 295–307 (2000).
23. Baron-Cohen S, Ring H, Moriarty J, Schmitz B, Costa D and Ell P. Recognition of mental state terms: clinical findings in children with autism and a functional neuroimaging study of normal adults. *Br J Psychiatry* **165**, 640–649 (1994).

Acknowledgements: This research was supported by a grant from the Office of the Vice President for Research of the University of Michigan and by a NSF Graduate Fellowship to D.L.