Reading Minds versus Following Rules:

Dissociating Theory of Mind and Executive Control in the Brain

Rebecca Saxe (1,2), Laura E Schulz (1) and Yuhong V Jiang (2)

(1) Department of Brain and Cognitive Sciences, MIT, Cambridge MA

(2) Psychology Department, Harvard University, Cambridge MA

Correspondence should be addressed to:

Rebecca Saxe

46-4019, 43 Vassar St, MIT

Cambridge, MA 02138

saxe@mit.edu

Acknowledgements:

Thanks to J. Gallant, R. Ivry, and N. Kanwisher for helpful comments and suggestions, and to C. Zhao for help collecting and analysing the data. The fMRI resources used in this study were supported by NCRR grant 41RR14075, the MIND Institute, and the Athinoula A. Martinos Center for Biomedical Imaging.

Running header: Reading minds versus following rules

Keywords: Theory of Mind, executive control, fMRI, temporo-parietal junction

Abstract

The false belief task commonly used in the study of Theory of Mind (TOM) requires participants to select among competing responses and inhibit prepotent responses, giving rise to three possibilities: 1) the false belief tasks might require only executive function abilities and there may be no domain-specific component; 2) executive control might be necessary for the emergence of TOM in development but play no role in adult mental state inferences; 3) executive control and domain-specific TOM abilities might both be implicated. We use fMRI in healthy adults to dissociate these possibilities. We find that non-overlapping brain regions were implicated selectively in response selection and belief attribution, that belief attribution tasks recruit brain regions associated with response selection as much as well-matched control tasks, and that regions associated with TOM (e.g. the right temporo-parietal junction) were implicated only in the belief attribution tasks. These results suggest that both domain-general and domain-specific cognitive resources are involved in adult TOM.

The development of social cognition – reasoning about people's thoughts and the mental causes of human behaviour – has been studied most intensely using a single task: the false belief task. In the basic design, a child watches while a puppet places an object in location A. The puppet leaves the scene and the object is transferred to location B. The puppet returns and the child is asked to predict where the puppet will look for the object. Three-year-olds think the puppet will look in location B, where the object actually is; older children think the puppet will look in location A, where the puppet last saw the object (Wimmer and Perner 1983, Wellman, Cross and Watson 2001). The standard interpretation of these results is that 3-year-olds lack a representational theory of mind. That is, 3-year-olds fail to understand how the contents of thoughts can differ from reality (Gopnik and Astington 1998; Gopnik and Wellman 1992).

As many researchers have noted (e.g. Roth and Leslie 1998; Perner and Lang 1999; Bloom and German 2002) though, children might fail the false belief task for reasons having nothing to do with deficits in understanding other minds. In particular, the false belief task requires a high level of executive control – that is, the ability to plan and carry out a sequence of thoughts or actions, while inhibiting distracting alternatives. Children must ignore a salient location (where the object is), attend to a non-salient location (where the object is not) and inhibit a highly practiced behaviour: pointing to the actual location of objects. Thus researchers have suggested that the false belief task (and a host of similar tasks) may underestimate children's knowledge about mental states (Bloom and German, 2002). Alleged shifts in children's theory of mind might reflect only changes in children's executive function – especially the abilities to select among competing responses, and to inhibit the prepotent one (Carlson, Moses and Claxton, 2004). (For present purposes, we will use "executive control" to refer collectively to the processes of response selection and inhibition).

Three independent lines of research suggest that response selection and inhibitory control play a non-trivial role in the false belief task. First, 3-year-olds are more likely to pass versions of the task that reduce the demands on inhibition (e.g Lewis and Osborne, 1990; Mitchell and Lacohee 1991; Wellman and Bartsch, 1998; Wellman, Cross and Watson 2001; Yazdi et al 2005). Thus for instance, children do better when they are allowed to use arrows to identify the object's location than when they must rely on practiced gestures like pointing. Second, young children who fail the false belief task also fail control tasks that place comparable demands on executive function but do not include any reference to other minds or mental states (Roth and Leslie, 1998; Slaughter, 1998; Zatichik, 1990). In a protocol closely matched to the false belief task, a photograph is taken of the object in location A, the object is moved to location B, and children are

asked where the object will be in the photo. No mental state understanding is required but 3-year-olds fail the task. Finally, individual differences in executive control are strongly correlated with individual differences in children's performance on theory of mind tasks (Carlson and Moses, 2001; Sabbagh, Xu, Carlon, Moses, and Lee, 2006).

Still, many developmental researchers have contested the centrality of executive control accounts. Researchers note that age-related deficits on the false belief task persist even when executive demands are substantially reduced (Wellman, Cross and Watson, 2001). Children under three fail even the simplest versions of the task (though see recent work by Onishi and Baillargeon, 2005). Also, group differences in executive control and Theory of Mind skills can be dissociated. For example, although performance on executive tasks is correlated in individuals in a single culture, Chinese preschoolers out-perform American children on every test of executive control, but show no advantage on standard Theory of Mind tasks (Sabbagh et al 2006). Executive control and Theory of Mind skills may also be dissociable in neuropsychological populations, such as children with autism. Although children with autism do show some executive function deficits (Ozonoff, Pennington and Rogers 1991; Frith 1997) and these deficits are correlated with performance on Theory of Mind tasks in individual subjects (Zelazo et al 2002), autistic children who fail false belief tasks nevertheless perform well on closely-matched false photograph control tasks (Leslie and Thaiss 1992; Charman and Baron-Cohen 1992).

Importantly, multiple distinct proposals for the relationship between executive function and theory of mind development have emerged (for more alternatives, see also Perner and Lang 1999; Moses 2001; McKinnon and Moscovitch 2006; German and Hehman 2005; Flynn, O'Malley and Wood 2004). According to the "Performance" hypothesis, young children fail false belief tasks not because of developmental changes in Theory of Mind, but because they are hampered by peripheral executive demands of the task. There are two possible versions of the Performance hypothesis. The lean version suggests that because performance changes on false belief tasks really reflect the development of executive function, there is no compelling evidence for any domain-specific component of Theory of Mind. The rich version of the Performance hypothesis, by contrast, proposes that delayed development of executive control masks three-year-olds' sophisticated representations of others' mental states. Alan Leslie and colleagues (Leslie and Polizzi 1998; Leslie, German and Polizzi 2005; German and Hehman 2006) for example, suggest that candidate belief contents are produced automatically by a domain-specific "Theory of Mind Mechanism," but these representations cannot be used to make behaviour predictions in the False Belief task until the children's executive skills become sufficiently robust.

Both lean and rich versions of the Performance hypothesis predict that executive control will continue to be necessary for false belief task performance even in adults, although only the rich version predicts that adults will additionally depend on domain-specific mechanisms for belief attribution but not other executive tasks. By contrast, the "Emergence" hypothesis suggests that executive control might be necessary for the emergence of representational theory of mind in development (i.e., it might be a pre-requisite for the construction of a concept of belief) but play no role in adult mental state inferences. In the current study, we use brain imaging to try to tease apart these alternatives.

In the past five years, the neural bases of human adult executive function and theory of mind have been investigated extensively, but separately. Executive control has multiple distinct components, each of which could contribute to belief attribution involved in theory of mind tasks. These components include monitoring and detecting conflict between competing representations or responses, selecting the correct response, and inhibiting the incorrect (possibly prepotent) response. Neuroimaging studies have attempted to distinguish the neural correlates of these components (e.g. Aron et al., 2004; Botvinick et al 1999; Konishi et al 1999; Braver et al 2001; Milham et al 2001; Garavan et al 2002; Sylvester 2003; Wager et al., 2005).

Across a range of tasks, including the Eriksen flanker task (Eriksen and Eriksen, 1974, Botvinick et al., 1999; Wager et al., 2005), the Stroop task (Stroop, 1938, Herd et al., 2006; MacDonald et al., 2000; Milham et al., 2001), and the Go/No-Go task (Braver et al., 2001; Wager et al., 2005), response conflict and inhibition of prepotent responses is strongly correlated with activity in anterior cingulate cortex (ACC), dorsolateral prefrontal cortex. superior parietal lobule, and ventralateral prefrontal cortex (Aron et al., 2004; Braver et al., 2001; Wager et al., 2005). Selecting the appropriate response while maintaining a high attentional load is associated with additional activity in bilateral frontal eye fields, intraprietal sulcus, and frontal operculum (Culham et al., 2001; Jiang & Kanwisher, 2003a, b, Jovicich et al., 2001).

Functional neuroimaging studies have also identified a set of brain regions putatively involved in reasoning about other people's minds, including right and left temporo-parietal junction, medial prefrontal cortex, posterior cingulate cortex, and the amygdala (for a recent review see Saxe 2006). Following tradition in developmental psychology, many of these previous studies have used false belief scenarios to elicit theory of mind reasoning, contrasted with control conditions that require physical or mechanical inferences (Fletcher et al., 1995, Goel et al 1995, Saxe & Kanwisher, 2003). Other studies required subjects to attribute communicative intentions (Walter et al 2004) or pretend mental states (German et al 2004). The response of at least one of

the implicated brain regions, the right temporo-parietal junction (RTPJ), is very selective (but see Decety and Grezes 2006); the response in the RTPJ is high when subjects read stories that describe a character's true or false beliefs, but low when they read stories containing other information about a character, including her appearance, cultural background, or even internal, subjective sensations - like being tired or achy or hungry (Saxe & Powell 2006, Saxe & Wexler 2005). For the most part, though, previous neuroimaging studies have aimed to match the executive demands of experimental and control conditions (e.g. Saxe and Kanwisher 2003), rather than directly investigate the relationship between executive function and Theory of Mind in the brain.

The current study used a two-tiered analysis to investigate the relationship between belief attribution and executive function in the adult brain. First, for reference, we replicated two previous experiments, one tapping components of executive function including response selection and inhibitory control (Jiang and Kanwisher 2003b) and the other requiring belief attribution (Saxe and Kanwisher 2003), within the same individual subjects (see Figure 1). Direct comparison within individual subjects allowed us to test hypotheses about the anatomical relationship between regions recruited by the two tasks (e.g. are these sets of regions truly non-overlapping in individual brains?) and the functional profiles of these regions.

Second, we investigated the recruitment of these same brain regions in a new, non-verbal version of the classic false belief task (called the "Chocolate-Box Task"). The stimuli were short animated films of a girl, a chocolate bar and two boxes. In each animation, the chocolate bar went into one box, the girl turned away, the chocolate bar came back to the middle; the girl either continued to face away or turned back to face the boxes; the chocolate either returned to the original box or moves to the new box (see Figure 2).

Participants completed the Chocolate-Box task under two different instruction sets. One set of task instructions (the *Algorithm* task) asked participants to treat the girl's final position as an arbitrary cue. The girl's final position "facing" the chocolates was a signal to indicate the final position of the chocolate; if the girl's final position was "away" from the chocolates, subjects were asked to indicate the first position of the chocolate. The other set of instructions (the *Theory of Mind* task) implemented a variant of the standard false belief task: participants were asked to identify "where the girl thinks chocolate bar is". For any combination of the girl's position and box location, these two rules generated identical responses (see Figure 2). To prevent hysteresis from the *Theory of Mind* instructions, subjects always performed the *Algorithm* task first. In the baseline condition, subjects simply identified the final position of the chocolate.

Relative to the baseline condition, success on the *Algorithm* task required subjects to monitor conflict between two competing responses, select the correct response, and (on half the trials) inhibit the prepotent tendency to respond based on the final position of the chocolate bar. Thus, we predicted that the *Algorithm* task would draw on the same neural resources as the "Response Selection" reference experiment.

On the other hand, the *Theory of Mind* instructions required the participant to make inferences about the girl's mental states, which the *Algorithm* task did not. We asked first, whether brain regions associated with belief attribution would be differentially recruited when subjects followed the *Theory of Mind* instructions; and second, whether performing the *Theory of Mind* task would also continue to recruit the brain regions associated with domain general executive functions.

Methods

Twelve naïve right-handed healthy adults (aged 18 – 26 years) participated for payment. All subjects were native speakers of English, and had normal or corrected-to-normal vision. Subjects gave written informed consent in accordance with the requirements of Internal Review Boards at Massachusetts General Hospital and MIT.

Subjects were scanned in a 3T scanner at the MGH facility in Charlestown, Massachusetts using twenty-six 4-mm-thick near-axial slices covering the whole brain except for the bottom of the cerebellum. Functional scans used TR = 2 s., TE=40 ms.

For the Chocolate-Box experiment, subjects participated in a half hour long practice session in the week before the scan. Each trial consisted of a short animation (see Figure 2). On the right half of the screen, a black rectangle labelled "chocolate" moved between two blue boxes located in the TOP and BOTTOM right hand corners. In each animation, the chocolate started in the center, moved into one of the randomly selected boxes (FIRST box), back to the center, and then moved into a box that might or might not be the same as the first box (LAST box). On the left hand side of the screen was a line-drawing of a girl. In each animation, the girl started FACING the boxes, then turned AWAY from the boxes. In half of the animations, the girl then turned back to FACING. There were thus four movie conditions: (same vs. different boxes for FIRST and LAST) x (girl FACING vs. girl AWAY). Two movies were constructed for each condition, one for each possible LAST box (TOP and BOTTOM).

Subjects learned to perform three different tasks. In each case, subjects responded to the position of the "chocolate," by pressing '1' for the TOP box, '2' for the BOTTOM box. The order in which subjects learned the response rules was counter-balanced across subjects.

Response rule 1 was: "Facing = last; Away first. If the girl is *facing* the boxes at the end of the trial, press the button for the *last* box. If the girl is *looking away from* the boxes, press the button for the *first* box." Response rule 2 was: "Facing = first; Away last. If the girl is *facing* the boxes at the end of the trial, press the button for the *first* box. If the girl is *looking away from* the boxes, press the button for the *last* box." Response rule 3 (the baseline condition) was: "Facing = last; Away last. If the girl is *facing* the boxes at the end of the trial, press the button for the *last* box. If the girl is *looking away from* the boxes, press the button for the *last* box." Rule 3 was included as a baseline condition, to control for the visual and motor components of the task.

During the practice session, subjects were instructed: "You will get a chance to practice these rules, first separately and then together. Try very hard to answer correctly on every trial. You will be told your accuracy score after each block. You will have 2 seconds after each movie to make your response, which is more than enough once you are used to the task." The practice session continued for approximately half an hour, until the subject was performing at above 85% correct on all three tasks.

Once in the scanner, subjects performed the tasks that they had practiced for three runs. Each run lasted 6 minutes and 52 seconds, and included six blocks, two of each task, presented in one of three orders: ABCCBA, CABBAC, or BCAACB, with 14 seconds of rest between blocks. Each block began with an instruction screen (e.g. "Rule 1: Facing = last; Away = First") presented for 4 seconds. Throughout the block, the rule number (e.g. "1") remained at the top of the screen. Each block then consisted of ten trials. Each trial included one animation (lasting 3 seconds) followed by a response prompt ("?") for 2 seconds.

After three runs, the subject (still in the scanner) was asked two simple debriefing questions. (A) "Did you notice that subjectively, rule 1 and rule 2 felt different, even though the instructions are symmetric?" (B) "Did you notice whether you were thinking about what the girl was thinking or seeing at any point in the experiment?"

Then subjects read a new task instruction: "Before you do three more runs of this task, here is something to think about. Rule 1 is: facing = last, away = first. Many subjects find this task easier if, instead of thinking of the rule, they ask themselves: "where does the girl think the chocolate is?" If she is looking away, she only saw it go into the *first* box. However, if she is

facing the boxes, then she saw it go into the *last* box. For the next three blocks, please try to use this new strategy instead of Rule 1. Please keep doing the same thing as ever for Rules 2 and 3, which cannot be solved by any other strategy. Is this fine and clear?" Subjects were then given a single practice example: "OK, just to check, here is an example. If the girl sees it go into the top box, and then turns away, and it goes into the bottom box, where does she think the chocolate is? She thinks it's in the bottom box? Or the top box?"

This interaction typically lasted about a minute. Then subjects immediately participated in three more runs of the experiment. The visual stimuli were exactly identical to the previous three runs (including the instruction screens). To prevent hysteresis from the easier *Theory of Mind* construal, subjects always participated in the two tasks in the same order. At the end of the session, each subject was asked two further questions: (A) Did you feel that you used the new instructions for Rule 1? And (B) Did you find the new instructions felt subjectively easier than the original version?

While subjects were in the scanner, they also participated in two reference experiments. In the "Theory of Mind" reference experiment, subjects read short verbal stories that required inferences about a character's beliefs with stories about a physical representation (e.g. photograph or map) that became outdated, and answered a fill-in-the-blank question about each story. Stimuli and story presentation were exactly as described in (Saxe and Kanwisher 2003; Experiment 2; see Figure 1 for an example). Stories in the two conditions contained the same number of words on average (Belief stories mean = 32 words; Photo stories mean = 32 words). Stories were presented for ten seconds, followed by four seconds for the question. The Belief and Photograph conditions alternated; the order of stories was counterbalanced across runs and across subjects.

In the "Response Selection" reference experiment, subjects performed a visual target selection task according to two rules that differed in stimulus-to-response compatibility. In both tasks, four lines were presented at four horizontally arrayed locations. Three of the lines were identical in length while the other one was different. Subjects were told to report the location of the unique line by pressing one of four horizontally arrayed keys. In the "compatible" condition (C), subjects pressed the key at the same location as the target; in the "incompatible" condition (I), subjects pressed a key that was two locations shifted from the target location. The two conditions were tested in separate blocks that lasted 48 seconds, in the order CIIC or ICCI. Stimuli and design were exactly as described in (Jiang and Kanwisher, 2003b, Experiment 1). The incompatible mapping condition involved suppression of a prepotent response (the tendency to press the key at the target's location) and mapping of stimulus onto an arbitrarily selected rule.

The comparison between incompatible and compatible mapping rules can reliably localize brain regions involved in central executive functions (Jiang & Kanwisher, 2003a.b; Schumacher et al., 2003).

fMRI Analysis

MRI data were analysed using SPM 99 (http://www.fil.ion.ucl.ac.uk/spm/spm99.html) and in-house software. Each subject's data were motion corrected and then normalised onto a common brain space (the MNI template). Data were then smoothed using a Gaussian filter (Full Width Half Maximum = 5 mm), and high-pass filtered during analysis. Every experiment used a blocked design, and was modelled using a boxcar regressor.

Following previous studies of theory of mind and executive function, regions of interest (ROI) were defined for each subject based on a whole brain analysis of the reference experiments, and defined as contiguous voxels that were significantly more active ($p<0.0001$, uncorrected) for the relevant contrast. Four regions were defined based on the 'Theory of Mind' reference experiment, using the contrast of stories about Beliefs versus Photographs: right and left temporo-parietal junctions (RTPJ and LTPJ), medial prefrontal cortex (MPFC) and posterior cingulate (PC). An additional nine regions of interest (ROIs) were defined based on the 'Response Selection' reference experiment, using the contrast of Incompatible versus Compatible Response mapping: right and left frontal eye fields (RFEF and LFEF), right and left intra-parietal sulcus (RIPS and LIPS), right and left middle frontal gyrus (RGFm and LGFm), right and left frontal operculum (RFO and LFO) and anterior cingulate cortex (ACC). Subcortical regions, such as thalamus and cerebellum, previously associated with response selection were not included; the present analyses were limited to the cortex.

The responses of these regions of interest were then measured while subjects performed the three tasks during the current experiment, separately for the first three runs (original instructions) and the last three runs (new instructions). Within the ROI, the average Percent Signal Change (PSC) relative to the control condition (PSC = 100 * raw BOLD magnitude for (condition – Control Task)/ raw BOLD magnitude for fixation) was calculated for each condition at each time point (averaging across all voxels in the ROI as well as all blocks of the same condition). These values were then entered into repeated measures ANOVAs. To avoid false negatives, all ROI analyses used an alpha level of 0.05. Because the data defining the ROIs were independent from the data used in the repeated measures statistics, the likelihood of false positives was drastically reduced.

Results

Behavioural results

Theory of Mind reference experiment: As in previous experiments (Saxe and Kanwisher 2003), subjects answered questions about False Photograph stories (mean 2.89 seconds) more slowly than questions about False Belief stories (mean 2.63 seconds, $t(11)=4.43$. $p<0.001$).

Response selection reference experiment: Due to technical difficulties, behavioural data from three subjects were lost. For the remaining subjects, as expected, subjects' reaction times were slower for the Incompatible response mapping (mean 707 ms) than for the Compatible mapping (mean 483 ms, $t(8) = 7.71$, $p < 0.0001$). Accuracy on the two response mappings were not significantly different (Incompatible = 90% correct; Compatible = 97% correct; $t(8) = 1.62$, $p > .14$). One subject was an outlier, performing at 55% accuracy in the Incompatible rule (chance is 25%). When this subject was excluded, the overall accuracy on the Incompatible response mapping was 97%.

Chocolate-Box experiment: In the brief debriefing session between the first half and second half of the experiment, no subjects reported spontaneously using a "Theory of Mind" construal to solve Rule One. After the change of instructions, 11/12 subjects reported feeling that they were using the Theory of Mind construal for Rule One, and that the new instructions felt easier than the original rule. One subject reported that the Theory of Mind rule was harder, and she did not consistently use it.

Reaction time costs relative to the control condition are shown in Figure 4 (top right corner). Behavioural data from a single subject in the first half of the experiment were lost due to technical errors. Reaction time data were analysed in two planned contrasts. First, we tested whether the *Algorithm* task incurred a reaction time cost, relative to the control task. Paired-samples t-tests revealed a highly significant increase in reaction time on the Algorithm version of Rule One ($t(10)=10.1$, $p<0.001$), and on Rule Two in the first half (($t(10)=9.6$, $p<0.001$) and the second half ($t(11)=5.9$, $p<0.001$).

Second, we investigated the behavioural correlate of switching from the *Algorithm* construal of Rule One to the *Theory of Mind* construal. To avoid confounding simple effects of practice and time, we tested for an interaction between Rule (One versus Two) and Time (first half [*Algorithm* rule] versus second half [*T.o.M* rule]). There was a main effect of experiment Half ($F(1,10)=9.3$, $p<0.05$), indicating that subjects responded faster overall on the second half of

the experiment; but critically, this main effect was driven by a significant interaction between rule and half ($F(1,10)=7.2$, $p<0.05$). There was no main effect of rule ($F(1,10)=0.5$, ns). To explore this interaction further, post-hoc t-tests were conducted on the difference between first and second half, separately for the two rules. The reaction times were significantly lower for Rule One for the *Theory of Mind* construal than for the *Algorithm* construal ($t(10)=3.5$, $p<0.01$); there was no effect of Time on reaction times for Rule Two ($t(10)=0.6$, ns).

fMRI results

In individual subjects, the two reference experiments, designed to replicate previous studies of Theory of Mind and Response Selection respectively, recruited the predicted, non-overlapping sets of brain regions (Table 1, Figure 3).

Based on the reference experiments, seven of the regions of interest could be identified in more than half of the individual subjects at the pre-established threshold of $p<0.0001$ (RTPJ, LTPJ, PC, RFEF, LFEF, RIPS and LIPS). The remaining six ROIs were not sufficiently reliable to identify in individual subjects. These regions were still of interest, though, because they had been identified in previous studies as regions associated with either Theory of Mind or response selection. So ROIs for these regions were defined based on the Random Effects analysis ($p<0.001$, uncorrected) of the group average data from the Belief vs Photo contrast (MPFC) or the Incompatible vs Compatible contrast (RGFm and LGFm, RFO and LFO, ACC). Peak coordinates for the group results are shown in Table 1. Average peak coordinates for the individual subject ROIs are shown in Table 2.

The response of each ROI to the two conditions in each of the reference tasks was then calculated. The ROIs identified based on the Response Selection task all showed a similar pattern of response during the Belief and Photo stories: relative to the two conditions in the "Response Selection" reference experiment, both conditions in the Theory of Mind experiment elicited an intermediate response, significantly higher than the Compatible control condition (with low response selection demands, all $t>2$, all $p<0.05$, paired-samples t-tests). There was no difference in the responses to the Belief and Photo stories (all $t<1$, all $p>0.2$, paired-samples t-test. See Figure 3). The ROIs identified based on the Theory of Mind stories task did not have a higher response during the Incompatible condition (high response selection) relative to the Compatible condition of the Response Selection. On the contrary, both the RTPJ and LTPJ showed a small effect in the opposite direction: the response in these regions was suppressed during the Incompatible condition (both $t>2$, $p<0.05$, paired-samples t-test, for a discussion of negative

activations see Gusnard and Raichle 2001; Mitchell 2006). These results complement and extend previous conclusions that domain-general Response Selection tasks, and the Theory of Mind stories task, recruit distinct sets of brain regions. No brain region identified based on one of the reference contrasts showed any hint of the effect in the other reference contrast. Domain-general response selection produced no increase in response in any brain region associated with Theory of Mind; in some of the ROIs, response selection demands suppressed the response. By contrast, both the Belief and Photo stories appeared to recruit resources of domain-general executive function to a significant, but similar, extent.

Next, we analysed the response of the same ROIs during the Chocolate-Box task. The questions of main interest were (1) which brain regions were recruited while subjects performed Rule One using the *Algorithm* construal, (2) which of these regions were also recruited when subjects used the *Theory of Mind* construal, and (3) which brain regions were recruited differentially just when subjects used the *Theory of Mind* construal? The response of the thirteen ROIs during the current experiment was measured in terms of percent signal change relative to the baseline condition ("Say which box the chocolate went into last"), which controlled for the visual and motor components of the experiment.

First, all of the regions of interest identified based on the "Response Selection" localiser experiment (Incongruent > Congruent responses) showed significantly higher BOLD signals during the *Algorithm* version of Rule One than during the control task (Figure 4). For the individually defined regions of interest, one-samples t-test of the percent signal change in Rule One, First Half were all highly significant (LIPS $t(8)=4.9$, $p<0.001$; RIPS $t(9)=4.5$, $p<0.005$; LFEF $t(7)=7.7$, $p<0.001$; RFEF $t(11)=5.5$, $p<0.001$). In the five regions identified from the group data (ACC, LFO, LGFM, RFO and RGFM), the same contrast was also highly significant (all $t(11)>4.5$, $p<0.001$).

By contrast, none of the regions identified based on the "Theory of Mind" localiser experiment (Belief > Photo stories) showed a higher response to the Algorithm version of Rule One than to the baseline condition. On the contrary, the LTPJ and RTPJ both showed trends in the opposite direction (RTPJ $t(10)=-1.9$, $p<0.1$; LTPJ $t(7)=-2.3$, $p<0.06$; similar numerical trends in the posterior cingulate and MPFC did not approach significance).

Second, we asked whether the brain regions showing a high response for the *Algorithm* construal of Rule One, would also be recruited when subjects switched to the *Theory of Mind* construal. That is, would performing a belief attribution task demand response selection resources? To control for general effects of time and practice, we analysed these results in a

repeated measures ANOVA, with two factors: Rule (One versus Two) and Time (first half versus Second half of the experiment). Since Rule Two had no change in construal, and could not be interpreted as a belief attribution task, Rule Two served as a control. A selective reduction in response selection demands for Rule One following the change in construal would lead to an interaction between Rule and Time. No such interaction was observed in any of the nine regions of interest that had shown a positive response during the *Algorithm* version (RIPS $F(1,9)=1.2$, $p>0.3$; LFEF $F(1,7)=1.6$, $p>0.2$; all others $F<0.05$, $p>0.8$). That is, performing the *Theory of Mind* task recruited the same "Response Selection" brain regions, and to the same degree, as did tasks tapping domain-general response selection and inhibitory control. In the Right and Left IPS, there were also no main effects of Rule or Time. The right and left FEF, and the ACC did show a main effect of time ($Fs>= 6$, $ps<0.05$); the response to both the *Theory of Mind* and control tasks were lower in the second half of the experiment than in the first half. (Similar effects in the Gfm and FO, bilaterally, did not reach significance.) Reductions of both the response in "response selection" brain regions and of reaction time in the second half of the experiment may reflect subjects' improvement with practice, both on the specific tasks and on task switching (Figure 4).

Third, we tested whether any of the remaining regions of interest (all of which were identified by the "Theory of Mind" localiser experiment) would be recruited differentially when subjects performed Rule One using the *Theory of Mind* construal. Again, to control for effects of time and practice, this hypothesis was tested using the same repeated measures ANOVA, with two factors: Rule (One versus Two) and Time (first half versus Second half of the experiment). One region, the RTPJ, did show the predicted interaction of Rule and Time ($F(1,10)=9.2$, $p<0.02$) and no main effects (Figure 4). That is, the response of the RTPJ was higher in Rule One during the *Theory of Mind* construal than in the *Algorithm* construal, but showed no difference with Time for Rule Two. In the posterior cingulate and LTPJ, the interaction approached significance (PC $F(1,11)=4.4$, $p<0.07$; LTPJ $F(1,7)=3.8$, $p<0.1$). In the MPFC ROI identified based on the group data, there were no significant effects or interactions. In all, the predicted pattern of domain-specific increased activation only for the Theory of Mind condition in the Chocolate-Box experiment was observed in one brain region, functionally defined in individual subjects based on the Theory of Mind reference task: the RTPJ.

Discussion

Belief attribution recruits both brain regions associated with domain-general attention, response selection and inhibitory control, and in addition at least one brain region associated with domain-specific representations of the contents of others' thoughts. The reference tasks compared

previously established tasks for executive control and Theory of Mind in the same individual subjects. Almost entirely non-overlapping brain regions were implicated in response selection (Incompatible vs Compatible response assignment) and in belief attribution (False Belief vs False Photograph stories). The results of the Chocolate-Box task support the same conclusions. Although they were given the same physical stimuli, and made the same correct responses, when subjects construed their task in terms of belief attribution, they responded faster, and selectively recruited an additional brain region than in the control task: the right temporo-parietal junction. Recruitment of brain regions implicated in executive control neither increased nor (in spite of reduced reaction times) decreased selectively for the belief attribution task.

The current results are consistent with an emerging consensus that executive control (including response selection and inhibition) is necessary for adult performance of at least some belief attribution (or "Theory of Mind") tasks, but that the construction of representations of others' thoughts relies on independent domain-specific cognitive and neural substrates. Apperly and colleagues (2004; Samson et al 2004; 2005), for example, reported a neuropsychological dissociation between these two contributions to belief attribution. WBA, a patient with right lateral frontal damage, was impaired on a range of executive control tasks, including belief attributions tasks with high inhibitory demands, but also including a range of control tasks that required stimulus or response selection (see also Fine, Lumsden and Blair, 2001). WBA was not impaired on a False Belief task when inhibitory demands were decreased (he did not know the real state of affairs and so did not have to inhibit his own true belief). By contrast, three patients with left temporo-parietal junction lesions were not impaired on general tests of inhibitory control, but nevertheless failed belief attribution tasks even when inhibitory demands were minimised.

The existence of domain-specific neural and cognitive resources for producing representations of belief contents might predict that this process would be relatively immune to interference from other, unrelated cognitive processes. Interestingly, there is some recent evidence consistent with this prediction. McKinnon and Moscovitch (2006) compared performance on belief attribution tasks, with or without interference from a second, unrelated task (an auditory 2-back detection task). When response selection demands were high (on second-order belief attributions, e.g. "Brad thinks that Ellen thinks that her team deserved to win the competition"), belief attribution was impaired by the addition of the unrelated task. On the hand, performance on first-order belief attributions (e.g. "Brad thinks that Ellen's team won the competition") was not impaired by the dual task, consistent with the idea that forming

representations of others' beliefs is relatively immune to competition with other cognitive demands. Similarly, Den Ouden et al (2005) had subjects make belief attributions in the scanner, with or without a secondary task (holding in mind the intention to press a button on seeing a red background). While belief attributions alone recruited the same set of regions typically associated with Theory of Mind tasks (e.g. TPJ, MPFC), the addition of the dual task did not lead to increased responses in these regions, but instead led to recruitment of a distinct set of regions, including right parietal regions similar to those recruited by the "Algorithm" construal of the current task. There was no cost on the belief attribution task associated with the addition of the dual task.

Overall, adult behavioural and neuroscientific data seem to be broadly compatible with the rich version of the Performance hypothesis (Leslie et al 2005; German and Hehman 2006). Of course, the current neuroimaging data from adult subjects cannot address one central tenet of Leslie's theory: namely, whether the "Theory of Mind" mechanism is fully functional in very young children, long before they pass False Belief tasks, but masked by under-developed executive control. Consistent with Leslie's hypothesis, though, we found that adult belief attribution (in both verbal and non-verbal tasks) recruits both brain regions associated with response selection, and domain-specific brain regions implicated selectively in belief attribution. The response properties of these two sets of regions bear considerable resemblance to the two components of Leslie's model: a domain-specific "Theory of Mind Mechanism" that forms representations of mental states, and a "Selection Processor" that selects among competing responses to make task performance possible.

Nevertheless, the current results may be troublesome for some of the details of Leslie's proposal. First, Leslie has suggested that the response selection demands of False Belief tasks might be recruiting a distinct cognitive resource from other, general response selection tasks (Friedman and Leslie, 2004). That is, Leslie has proposed the existence of a domain-specific "Selection Processor" just for selecting among mental state representations. However, our findings suggest the opposite: belief attribution tasks recruit exactly the same brain regions as a whole range of (non-belief-related) response selection tasks. Also consistent with the current claim, recent studies of healthy older adults find that aging does not impair the ability to represent the contents of beliefs per se, but does selectively impact the subjects' ability to perform belief attribution tasks with high response selection or inhibitory demands (McKinnon and Moscovitch 2006; German and Hehman 2005). German and Hehman (2006) found that older adults' were disproportionately impaired on double-negated versions of belief attribution tasks (like predicting

what Sally will do if she doesn't want to hear music, and doesn't know that the band is playing at the coffee shop). This impairment was best explained by older adults' difficulties on domain-general tests of inhibitory control, like the Stroop task.

Second, Leslie has also proposed that the TOMM represents the possible contents of others' mental states automatically and obligatorily. If we identify Leslie's domain-specific TOMM with the belief-attribution-specific brain regions in the current paper, though, this proposal is in trouble (see also Apperly et al, In Press). The response of these brain regions is not an automatic or obligatory response to a given stimulus (e.g. the animated girl's position and motions), but depends on the subjects' intentional construal of that stimulus.

Some alternative hypotheses about the developmental relationship between executive function and Theory of Mind fare much worse in light of recent evidence, though. As described above, there is considerable reason to believe that belief attribution task performance relies on executive control in adults, ruling out the strong "Emergence" hypothesis that executive control is only necessary during the initial development of Theory of Mind (Moses 2001; see also Sabbagh et al 2006; although these authors may intend a weaker version of the "Emergence" hypothesis). Neural evidence for a set of brain regions recruited just for Theory of Mind tasks is also, prima facie, inconsistent with the lean version of the "Performance" hypothesis; namely, that there is no domain-specific component of Theory of Mind. Most clearly ruled out by the current data are some early claims that Theory of Mind and executive control are related in development because they rely on the same brain regions (Ozonoff, Pennington and Rogers 1991; Perner and Lang 1999; Ruby and Decety 2003). On the contrary, we found a striking lack of overlap in the brain regions implicated in executive control (specifically response selection and inhibition) and in Theory of Mind tasks, either in the separate reference experiments or within the Chocolate-Box experiment.

References

Apperly IA, Samson D, Chiavarino C, Humphreys GW. (2004) Frontal and temporo-parietal lobe contributions to theory of mind: neuropsychological evidence from a false-belief task with reduced language and executive demands. *J Cogn Neurosci*. 16(10):1773-84.

Aron AR, Robbins TW, Poldrack RA (2004). Inhibition and the right inferior frontal cortex. Trends in Cognitive Sciences, 8(4), 170-177.

Bloom P, German TP. (2000) Two reasons to abandon the false belief task as a test of theory of mind. *Cognition*. 77(1):B25-31.

Carlson SM, Moses LJ, Claxton LJ. (2004) Individual differences in executive functioning and theory of mind: An investigation of inhibitory control and planning ability. *J Exp Child Psychol*. 87(4):299-319.

Carlson SM, Moses LJ, Hix HR. (1998) The role of inhibitory processes in young children's difficulties with deception and false belief. *Child Dev*. 69(3):672-91.

Carlson SM, Moses LJ. (2001) Individual differences in inhibitory control and children's theory of mind. *Child Dev*. 2001 Jul-Aug;72(4):1032-53.

Charman T, Baron-Cohen S. (1992) Understanding drawings and beliefs: a further test of the metarepresentation theory of autism: a research note. *J Child Psychol Psychiatry*. 33(6):1105-12.

Culham JC, Cavanagh P, Kanwisher NG (2001). Attention response functions: Characterizing brain areas using fMRI activation during parametric variations of attentional load. *Neuron* 32(4), 737-745.

Decety J, Grezes J (2006) The power of simulation: imagining one's own and other's behavior. *Brain Research* 1079(1), 4-14.

den Ouden HE, Frith U, Frith C, Blakemore SJ. (2005) Thinking about intentions. *Neuroimage*. 28(4):787-96

Fine C, Lumsden J, Blair RJ. (2001) Dissociation between 'theory of mind' and executive functions in a patient with early left amygdala damage. Brain. 124(Pt 2):287-98.

Fletcher PC, Happé F, Frith U, Backer SC Dolan, RJ. (1995). Other minds in the brain: A functional imaging study of ‘theory of mind’ in story comprehension. *Cognition, 57*, 109-128.

Flynn E, O'Malley C, Wood D. (2004) A longitudinal, microgenetic study of the emergence of false belief understanding and inhibition skills. *Dev Sci*. 7(1):103-15.

Friedman O, Leslie AM (2004) A developmental shift in processes underlying successful belief-desire reasoning. *Cognitive Science*, 28(6):963-77

Frith U (1997) The neurocognitive basis of autism. *Trends Cog Sci* 1:73-7

German TP, Hehman JA. (In Press) Representational and executive selection resources in 'theory of mind': Evidence from compromised belief-desire reasoning in old age. *Cognition*.

German TP, Niehaus JL, Roarty MP, Giesbrecht B, Miller MB. (2004). Neural correlates of detecting pretense: Automatic engagement of the intentional stance under covert conditions. *Journal of Cognitive Neuroscience, 16*, 1805-1817.

Gopnik A, Wellman HM. (1992) Why the child's theory of mind really is a theory. *Mind and Language* 7 (1-2), 145-171.

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

Gusnard DA, Raichle ME (2001) Searching for a baseline: functional imaging and the resting human brain. *Nature Reviews Neuroscience*. 2, 685-94.

Herd SA, Banich MT, O'Reilly RC (2006). Neural mechanisms of cognitive control: An integrative model of stroop task performance and fMRI data. *Journal of Cognitive Neuroscience*, 18(1), 22-32.

Jiang Y, Kanwisher N. (2003a) Common neural mechanisms for response selection and perceptual processing. *J Cogn Neurosci*. 15(8), 1095-110.

Jiang Y, Kanwisher N. (2003b) Common neural substrates for response selection across modalities and mapping paradigms. *J Cogn Neurosci*. 15(8), 1080-94.

Jovicich J, Peters RJ, Koch C, Braun J, Chang L, Ernst T (2001). Brain areas specific for attentional load in a motion-tracking task. *J Cogn Neuro*, 13(8), 1048-1058.

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

Leslie A, Thaiss L. (1992). Domain specificity in conceptual development. *Cognition* 43: 225-51

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

Leslie AM, Friedman O, German TP. (2004) Core mechanisms in "theory of mind". *Trends Cogn Sci*. 2004 Dec;8(12):528-33.

Leslie AM, German TP, Polizzi P. (2005) Belief-desire reasoning as a process of selection. *Cognit Psychol*. 50(1):45-85.

Lewis C, Osborne A. (1990) Three-year-olds' problems with false belief: conceptual deficit or linguistic artifact? *Child Dev*. 61(5):1514-9

MacDonald AW 3rd, Cohen JD, Stenger VA, Carter CS (2000). Dissociating the role of the dorsolateral prefrontal and anterior cingulate cortex in cognitive control. *Science*, 288, 1835-1838.

McKinnon MC, Moscovitch M. (In Press) Domain-general contributions to social reasoning: Theory of mind and deontic reasoning re-explored. *Cognition*.

Mitchell JP (2006) Mentalizing and Marr: An information processing approach to the study of social cognition. *Brain Research* 1079, 66-75.

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

Onishi KH, Baillargeon R. (2005) Do 15-month-old infants understand false beliefs? *Science*. 308(5719):255-8.

Ozonoff S, Pennington BF, Rogers SJ (1991) Executive function deficits in high-functioning autistic children: relationship to Theory of Mind. *J. Child Psychol. Psychiatry* 32 1081-105

Perner J, Lang B, Kloo D. (2002) Theory of mind and self-control: more than a common problem of inhibition. *Child Dev*. 73(3):752-67.

Perner J, Lang B. (1999) Development of theory of Mind and executive control. *TICS* 3(9):337-344

Roth D, Leslie AM. (1998). Solving belief problems: Toward a task analysis. *Cognition*, 66, 1–31

Ruby P, Decety J. (2003) What you believe versus what you think they believe: a neuroimaging study of conceptual perspective-taking. *Eur J Neurosci.* 17(11):2475-80.

Sabbagh MA, Xu F, Carlson SM, Moses LJ, Lee K. (2006) The development of executive functioning and theory of mind. *Psychol Sci*. 17(1):74-81.

Samson D, Apperly IA, Kathirgamanathan U, Humphreys GW. (2005) Seeing it my way: a case of a selective deficit in inhibiting self-perspective.Brain. 128(Pt 5):1102-11

Samson, D., Apperly, I. A., Chiavarino, C., & Humphreys, G. W. (2004). The left tempero-parietal junction is necessary for representing someone else's belief. *Nature Neuroscience, 7*, 449-500.

Saxe R (2006) Uniquely human social cognition. *Current Opinion in Neurobiology*. 16:235–239

Saxe R, Powell L. (2006). It's the thought that counts: specific brain regions for one component of Theory of Mind. *Psychological Science* 17(8), 692-9.

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

Saxe, R. Wexler, A. (2005). Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia. 43*, 1391-9.

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

Schumacher EH, Elston PA, D'Esposito M. (2003) Neural evidence for representation-specific response selection. *J Cogn Neurosci*. 15(8):1111-21

Slaughter V. (1998) Children's understanding of pictorial and mental representations. *Child Dev*. 69(2):321-32.

Wager TD, Sylvester CY, Lacey SC, Nee DE, Franklin M, Jonides J (2005). Common and unique components of response inhibition revealed by fMRI. NeuroImage, 27(2), 323-340.

Walter H, Adenzato M, Ciaramidaro A, Enrici I, Pia L, Bara BG. (2004). Understanding Intentions in Social Interaction: The role of the Anterior Paracingulate Cortex. *J Cogn Neurosci.* 16(10):1854-63

Wellman HM, Bartsch K. (1998) Young children's reasoning about beliefs. *Cognition*. 30(3):239-77.

Wellman HM, Cross D. (2001). Theory of mind and conceptual change. *Child Development, 72*, 702-7.

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

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

Yazdi AA, German TP, Defeyter MA, Siegal M. (2005) Competence and performance in belief-desire reasoning across two cultures: The truth, the whole truth and nothing but the truth about false belief? *Cognition*.

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

Zelazo PD, Jacques S, Burack J, Frye D. (2002). The relation between theory of mind and rule use: Evidence from persons with autism-spectrum disorders. *Infant and Child Development (Special Issue: Executive function and its development*) 11:171-195

Tables

Table 1. Peak activations in the reference Experiments, Random Effects analysis, n=12, thresholds: p<0.001 uncorrected for single voxels, and p<0.05 for clusters.

Activated Region	**MNI coordinate**	**T value at the peak voxel**
Response Selection reference experiment		
Cortical regions		
Left intraparietal sulcus	-36 -51 51	10.42
Right intraparietal sulcus	36 -36 45	7.72
Pre-supplementary motor area/Anterior cingulate	0 12 57	8.62
Left Frontal Eye fields	-21 3 60	7.72
Right Frontal Eye fields	27 -6 54	6.04
Left frontal operculum	-30 24 6	10.94
Right frontal operculum	39 18 6	6.51
Left middle frontal gyrus	-42 27 24	7.37
Right middle frontal gyrus	51 15 33	7.54
Right middle temporal gyrus	51 -51 -3	7.27
Subcortical regions		
Left Thalamus	-6 -21 9	11.40
Right Thalamus	6 -18 9	8.90
Left midbrain	-6 -27 -15	8.16
Right midbrain	6 -27 -15	7.73
Theory of Mind reference experiment		
Cortical regions		
Posterior cingulate	0 –60 30	13.31
Right temporo-parietal junction	57 –57 18	12.48
Medial prefrontal cortex	0 57 12	8.28
Right anterior superior temporal sulcus	54 –18 -21	7.39
Left insula	-54 –12 12	7.10
Left temporo-parietal junction	-45 –72 39	6.52
Subcortical regions		
Right amygdala	24 3 -12	8.48
Left amygdala	-24 –6 -24	6.41

Table 2. Average peak voxels in the individual subjects' regions of interest, $p<0.0001$ uncorrected. Peaks are given in MNI coordinates.

Region of Interest	**Average Peak (Standard Dev)**	**Number of Subjects**
Response selection experiment		
Right Intraparietal sulcus	38(8) -55(10) 52 (10)	10
Left Intraparietal sulcus	-40(6) -50(6) 54(9)	10
Right Frontal Eye Fields	32(4) 2(10) 60(7)	10
Left Frontal Eye Fields	-32(5) 1(5) 61(5)	9
Theory of Mind experiment		
Right Temporo-parietal junction	54(6) -60(5) 18(7)	11
Left Temporo-parietal junction	-47(5) -65(6) 26(10)	8
Posterior Cingulate	2(6) -60(7) 32(8)	12

Figure Legends

1. **Stimuli and Tasks for the Reference Experiments**. Left-side panel: the "Response Selection" task. Subjects press a button to indicate one line that is different in length from the other three. Grey arrows show the response assignments in the Compatible and Incompatible conditions. The dark grey finger is the correct answer; the light grey finger is the "pre-potent" response, based on spatial alignment between the stimulus and the fingers. Right-side panel: the "Theory of Mind" task. Subjects first read a short verbal passage, and then responded to a fill-in-the-blank question via a button press. Subjects were not told about the division of the stories into the two conditions, "Belief" and "Photograph". In the examples shown, the dark grey finger is the correct answer; the light grey finger is the "pre-potent" response, based on reality (or the subject's own beliefs).

2. **Stimuli and Tasks for the Main Experiment.** Shown are single frames from two of the eight animations. Time goes from top to bottom. Also shown are the instructions for Rule One, in both the "Algorithm" and "Theory of Mind" construal, showing that the two construals generate the same responses for each animation.

3. **Brain responses in the Reference Experiments.** The top panel shows the rendered brain response in four individual subjects ($p<0.001$ uncorrected), on a standard anatomical brain. Shown in red are regions responding more during "Belief" than "Photograph" stories. Shown in green are regions responding more during "Incompatible" than "Compatible" blocks of the response selection task. Regions that show both of these profiles are shown in yellow; there are almost no such regions in any individual subject. The left-most brain shows the approximate location of eleven regions of interest, coloured to match the reference experiment from which they were derived: (1) Left GFm, (2) MPFC, (3) Right Gfm, (4) ACC, (5) Left FEF, (6) Right FEF, (7) Left IPS, (8) PC, (9) Right IPS, (10) Left TPJ, (11) Right TPJ. The botton panel shows the percent signal change in the BOLD response, relative to passive fixation, during Belief stories (Red), Photograph stories (Grey), Compatible (light green) and Incompatible (Dark Green) response selection, in seven individually defined regions of interest: (from left to right) Left FEF, Right FEF, Left IPS, Right IPS, PC, Left TPJ and Right TPJ.

4. **Brain responses in the Main Experiment.** The brain in the centre shows the results of Random Effects analysis of the two reference experiments: the response selection task in green, and the belief task in red (each $p<0.0001$ uncorrected). The graphs show the percent signal change relative to the baseline condition during the Chocolate-Box experiment, in thirteen regions of interest (ROIs), for two construals of Rule One ('Algorithm' = light blue, 'Theory of Mind' = dark blue) and for Rule Two in both halves of the experiment (first half = light grey, second half = dark grey). A selective response to 'Theory of Mind' is shown by a higher dark blue bar than any of the other three bars (in the TPJ only). In the top right hand corner are the reaction time costs for the same four conditions. A selective advantage for Theory of Mind instructions is shown by a lower dark blue bar than any of the other three bars. ROIs that were derived from the group analysis of the reference experiments, because the responses in the individuals were not sufficiently reliable, are marked with an asterix.