OXFORD

*Social Cognitive and Affective Neuroscience*, 2017, 1–9

doi: 10.1093/scan/nsx034
Original Manuscript

# Higher order intentionality tasks are cognitively more demanding

Penelope A. Lewis,[1] Amy Birch,[2] Alexander Hall,[1] and Robin I. M. Dunbar[3]

[1]School of Psychological Sciences, University of Manchester, Manchester, UK, [2]Division of Brain Sciences, Imperial College London, London, UK, and [3]Department of Experimental Psychology, University of Oxford, Oxford, UK

Correspondence should be addressed to Robin I. M. Dunbar, Department of Experimental Psychology, University of Oxford, South Parks Rd, Oxford OX1 3UD, UK. E-mail: robin.dunbar@psy.ox.ac.uk.

## Abstract

A central assumption that underpins much of the discussion of the role played by social cognition in brain evolution is that social cognition is unusually cognitively demanding. This assumption has never been tested. Here, we use a task in which participants read stories and then answered questions about the stories in a behavioural experiment (39 participants) and an fMRI experiment (17 participants) to show that mentalising requires more time for responses than factual memory of a matched complexity and also that higher orders of mentalising are disproportionately more demanding and require the recruitment of more neurons in brain regions known to be associated with theory of mind, including insula, posterior STS, temporal pole and cerebellum. These results have significant implications both for models of brain function and for models of brain evolution.

**Key words:** fMRI; mentalising; intentionality; reaction time; social brain

## Introduction

Mentalising, also known as mindreading or theory of mind, is the ability to infer the mental states of another individual and to recognise that these mental states can affect their behaviour (Premack and Woodruff, 1978). It is a trait that appears to be all but unique to humans (Saxe, 2006; Tomasello and Call, 1998). Formally, mentalising involves the recursive understanding of mental states (*I think that you suppose that I intend that you believe that something is the case* ...) and the number of separate mind states involved is defined as the order of intentionality (in this example, fourth order intentionality) (Dennett, 1983). While children begin to engage with others' mindstates as early as 12-18 months of age (Baillargeon *et al.*, 2010; Kovács *et al.*, 2010), formal theory of mind (i.e. when they understand the mind states of another person sufficiently well to recognise a false belief—equated with second order intentionality: *I believe that you think that something is the case* [*even when I know this isn't true*])—probably does not finally consolidate until around 4–5 years of age (Perner, 1991). Most research over the past two decades or so has focussed on formal theory of mind, i.e. second order intentionality, largely because this is a major developmental milestone for young children. However, as children develop, they are able to cope with progressively higher orders of mentalising (Henzi *et al.*, 2007), and are able to handle the mind states of several individuals at the same time (*I think that Peter believes that Susan wants Elizabeth to suppose* [*something*]). In normal adults, this capacity reaches an asymptotic limit at around fifth order intentionality, with only small numbers of individuals able to perform successfully at higher orders (Kinderman *et al.*, 1998; Stiller and Dunbar, 2007; Powell *et al.*, 2010).

Although Roth and Leslie's assertion (1998) that we have very little idea as to what, in cognitive terms, theory of mind and its associated higher orders actually is remains largely unchallenged, two key claims have been made about it: first, that this form of social cognition is computationally demanding (Dunbar, 1998; Lin *et al.*, 2010) and, second, that the ability to

**Received:** 27 May 2016; **Revised:** 16 January 2017; **Accepted:** 6 March 2017

engage in the higher orders of intentionality may be dependent on the capacity to recruit a more extended neural network (Barrett *et al*., 2003; Dunbar, 2010). Several studies have used reaction time paradigms to demonstrate that introspection is effortful (Corallo *et al*., 2008). However, the question of whether mentalising is any more demanding than the more conventional concatenation of a set of facts remains unclear. While two recent studies (one using memory for friends' traits as a 'social working memory' task (Meyer *et al*., 2012), the other using eye gaze in an implicit false belief task: Schneider *et al*., 2012) provide *prima facie* evidence for an effect of cognitive load, mentalising itself has yet to be examined.

Functional MRI has revealed a pattern of activation within certain regions of the brain that has been interpreted as a neural network for theory of mind reasoning. These include the temporo-parietal junction (TPJ), temporal pole (TP), and medial prefrontal cortex (mPFC) [20-24], as well as anterior cingulate cortex (ACC), superior temporal sulcus (STS), superior temporal gyrus (STG), precuneus/posterior cingulate cortex and amygdala (Gallagher *et al*., 2002; Frith and Frith, 2003; Gallagher and Frith, 2003; Fukui *et al*., 2006; Kobayashi *et al*., 2006; Vollm *et al*., 2006; Saxe and Powell, 2006; van Overwalle, 2009; Carrington and Bailey, 2009; Lewis *et al*., 2011). These studies do not, however, consider the pattern of activation as a function of the complexity (or level) of mentalising, but rather focus exclusively on theory of mind reasoning (i.e. second order intentionality).

Here, we first use a reaction time task to test the hypotheses (1) that mentalising (memory for mental states) is cognitively more demanding than simple memory for facts (with no mentalising component), and (2) that this difference is amplified at higher orders of task difficulty. We then use functional MRI (fMRI) to show that this involves a parallel increase in neural activation as the intentionality order of the task increases. In both studies, the factual tasks function as the baseline for determining the added effect due to mentalising. Both mentalising and factual tasks require short term memory for their successful completion, and the question we ask is whether or not adding a mentalising component to factual tasks increases the cognitive work that has to be done to answer them correctly.

## Materials and methods

To examine the impact of increasingly complex mentalising on both response times and neural activity, we used a vignette-plus-questions study design (Stiller and Dunbar, 2007; Lewis *et al*., 2011) in which subjects read a short story describing a social event and then answer a series of true/false questions at varying levels of intentionality, with control questions which required equivalent factual memory processing but contained no mentalising component. In both experiments, subjects read a series of short stories about different social situations that involved two or more individuals (for example, a man taking his wife out to dinner for their anniversary). The characters within the stories have different points of view about their fellow characters and the situation. The stories also contain facts about the social situation and the characters themselves. The stories presented were those used by Lewis *et al.* (2011), with a sixth newly written for Experiment 1. All of the stories were approximately 200 words long (mean length = 197 ± 12.6 se words). After reading each story, participants were presented with a series of statements relating to it, each of which could be true or false. Ten of these statements involved mentalising (two each at intentionality levels 2–6) and 10 were purely factual (with content and length matched, as well as the number of factual propositions involved) (with level 1 in each case representing the subject's own state of mind: *the subject believes that* ….). Mentalising questions did not differ significantly from factual questions in the number of words or propositions they contained (on average, mentalising and factual questions differed by no more than 0.5 words at any given level: *t*-tests, all levels NS). A sample story with questions is given in the Supplementary Data.

Stimuli were presented using Cogent v.2000 run through Matlab 6.5 software platform. Subjects were instructed to answer each question using a keyboard (with keys for 'true', 'false' and 'don't know') as soon as possible after the question was presented, and that the next question would not be presented until they had done so. The 'don't know' response option was included in order to minimise the incidence of guessing in the other options, thus ensuring a cleaner sample of trials for fMRI analysis. The timings of stimuli presentation were the same for all subjects, but the speed with which successive questions were asked depended on how fast individual subjects responded.

In line with institutional requirements in force at the time of the experiments, participants were volunteers and were not provided with any financial or other compensation.

### Experiment 1

Thirty-nine subjects (19 females; mean age 35.5 years, range 18–60) read six stories presented visually, with the text broken down into a series of 4–5 screens which participants viewed sequentially. All participants were healthy with no history of neurological or psychiatric disorders. After viewing each story, participants were presented sequentially with 20 True/False statements in random order. Reaction times and response accuracy were recorded. Only reaction times to correct responses are analysed.

### Experiment 2

For the fMRI experiment itself, 17 subjects (mean age 22 ± 2.9 years, 9 females) were presented with five of the stories seen by subjects in Experiment 1, following the design in Figure 1. All participants were healthy and right handed with no history of neurological or psychiatric disorders. Although participants were located within the MRI scanner throughout this experiment, scans were only collected while they answered the True/False questions. Note that for this experiment only questions at levels 2–4 are used so as to ensure that the task is well within the competences of normal adults who have an average competence at fifth order. Visual stimuli were presented on a projector screen viewed through a mirror attached to the head coil above the participant's head. The text of each story was broken into 4–5 separate screens, as in Experiment 1. Participants also heard each story being read through MRI compatible headphones while viewing the text (~1 min per story). After presentation of each story, subjects viewed 20 True/False statements relating to that story randomly intermixed with five null events (which comprised fixation alone, and were included to provide a baseline for the fMRI). Each statement was shown for a random duration between 7 and 11 seconds (based on the reaction time data from the pilot experiment, see ESM section B). After each statement, 'T F DK' [true, false, don't know] was displayed on the screen for 1.5 s as a cue to respond with the appropriate finger. The order from left to right of these cues was randomised with respect to the questions. There was a gap of 2–5 s (mean

3 s) blank screen between questions to provide jitter. In total, each session (the time during which the participants were being scanned) was 304 s long. Stories were presented in random order. Prior to scanning, but whilst already in the scanner, subjects read a practice story and answered ten questions using this same paradigm to ensure that they understood the task.

### Subject performance

Performance was calculated as percentage of correct responses in the task (don't know answers being classed as incorrect) on questions at each order of intentionality or equivalent factual memory. Although both intentionality and factual memory performance differed significantly from a normal distribution in this sample (Kolmogorov-Smirnov test with Lillefors Significance Correction: intentionality, $P = 0.01$; factual, $P = 0.058$; $N = 51$ in both cases), we used ANOVA for statistical analysis (i) because ANOVA is robust to departures from normality and (ii) to maintain consistency with the other analyses.

### MRI scanning

T2-weighted echo planar images (EPI) with BOLD (blood-oxygen-level-dependent) contrast were acquired using a specialized sequence which minimized signal dropout in the medial temporal lobe (Deichmann *et al.*, 2003). We used the following scanning parameters to achieve whole brain coverage: 50 oblique axial slices at a 20 degree tilt in the anterior-posterior axis, TR of 3 s, slice thickness of 2 mm (40% gap), TE of 30 ms, in plane resolution was 3 × 3 mm. Data were collected in five separate sessions (runs). Each run lasted 303 s, giving 101 volumes. High resolution anatomical whole brain images were obtained using a T1-weighted 3D-gradient-echo pulse sequence, with the following parameters: (T1 190°, TR 7.92 s, TE 2.48 ms, FOV 224 × 256, matrix 256 × 256 × 256 pixels, flip angle 16°), acquired in sagittal plane.

### fMRI analysis

Functional MRI images were analysed using the statistical parametric mapping (SPM2) software package (Wellcome Trust Centre for Neuroimaging London, UK, http://www.fil.ion.ucl.ac.uk/spm). After the first two volumes of each session were discarded to allow for T1 equilibration effects images were corrected for head motion by realigning with the first image of the first session and spatially normalised to an EPI template corresponding to the Montreal Neurological Institute (MNI) brain. Normalised images were smoothed using a Gaussian Kernel size with a full width at half-maximum (FWHM) of 8 mm.

To characterise functional responses, data were examined using a two-level random effects analysis. At the first level, the event-related design matrix contained all five of the experimental sessions (one for each story). The design matrix included four primary regressors for each of these sessions: *presentation time (mentalising)*, *presentation time (factual)*, *response time (mentalising)* and *response time (factual)* with both the response time regressors having three parametric regressors each: *button pressed*, *accuracy* and *level*. The four primary regressors were measured in seconds, the parametric regressors each had three options: *button pressed (2, 4, 8)*, *accuracy (true, false, don't know)* and *order* ($2^{nd}$, $3^{rd}$, $4^{th}$). Parameter estimates reflecting the height of the hemodynamic response function for each regressor were calculated at each voxel. Contrast images relating to specific combinations of correctly classified items were calculated. These included (i) mentalising, (ii) mentalising parametric modulation, (iii) factual and (iv) factual parametric modulation.

Next, the contrast images resulting from our first-level analysis were entered into two separate second level design matrixes in order to conduct two one-way ANOVAs. The first design matrix, containing the contrast images for mentalising and factual questions was used to conduct contrast 1 which examined the responses to the mentalising component of our task while controlling for memory (mentalising > factual). The second design matrix, containing contrast images from the parametric modulation of mentalising and factual, was used to conduct contrasts 2 which compared parametric modulation of mentalising and factual processing (parametric mentalising > parametric factual). For completeness, we also isolated areas associated with parametric modulation of mentalising (contrast 3) and factual memory (contrast 4) in isolation. In order to determine areas that were common to contrast 1 and contrast 2, both contrasts were plotted on the same brain in xjView and common areas (regions of overlap) were extracted. Statistical thresholding of the second-level activation maps associated with these contrasts was an uncorrected threshold of $P < 0.001$ in combination with a minimal cluster extent of 38 voxels. This yields a whole-brain alpha of $P < 0.05$, determined using a Monte-Carlo Simulation with 1000 iterations, using a function implemented in Matlab. Thresholded data were rendered onto the MNI canonical brain for visualisation ($P < 0.001$, $k > 38$). Areas which were parametrically modulated by mentalising but not factual and which also responded significantly more to the mentalising task than the factual task (mentalising > factual) were determined by plotting both results on the same brain.

### Data

The data for the reaction time experiments can be found in the ESM [S1 Experiments 1 + 2 Reaction time experiment data]. The neuroimaging data can be sourced at: 10.15127/1.269726

### Ethics

The specific study designs were approved by the respective Ethics Committees at Manchester and Liverpool Universities.

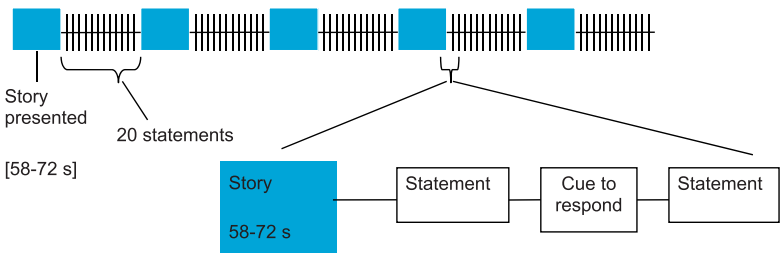


**Fig. 1.** The basic design for Experiments 1 and 2.

Subjects gave written informed consent on arriving at the imaging facilities.

## Results

We evaluate first the reaction time data from Experiments 1 and 2 to determine whether mentalising questions were more demanding (indexed as time taken to respond on correctly answered questions) than pure factual recall questions, and how this difference related to order of complexity. We then evaluate the functional results of the imaging experiment to determine whether there are any neurophysiological correlates.

### Reaction time tasks

Experiment 1 was the reaction time experiment, intended, first, to determine whether mentalising tasks take longer (i.e. are harder) to process than the factual memory tasks and, second, to ascertain whether this difference was amplified at higher levels of intentionality. Mean accuracy across the five question levels (levels 2–6) was 82.5% ± 1.1 se for mentalising tasks and 76.8% ± 1.1 se for factual tasks. The two types of task did not differ in the accuracy of responses: controlling for individual differences between subjects with question type and level as fixed factors there was, as might be expected, a significant effect of level (ANOVA: $F_{4,380} = 36.6$, $P < 0.001$, $\eta^2 = 0.215$) but not of question type (mentalising *vs* factual: $F_{1,380} = 0.05$, $P = 0.816$, $\eta^2 = 0.000$), with a significant interaction ($F_{4,380} = 18.64$, $P < 0.001$, $\eta^2 = 0.164$). In the case of reaction times for correctly answered questions only, by contrast, there was a significant difference between the mentalising and factual questions, and a significant effect due to level (Figure 2, $\text{mean}_{\text{mentalising}} = 6733 \pm 228.6$ se msecs *vs* $\text{mean}_{\text{factual}}$=5002 ± 143.7 se msecs; ANOVA, question type, $F_{1,379} = 103.8$, $P < 0.001$, $\eta^2 = 0.215$; level, $F_{4,379} = 134.7$, $P < 0.001$, $\eta^2 = 0.584$), with a significant interaction ($F_{4,379} = 14.7$, $P < 0.001$, $\eta^2 = 0.134$).

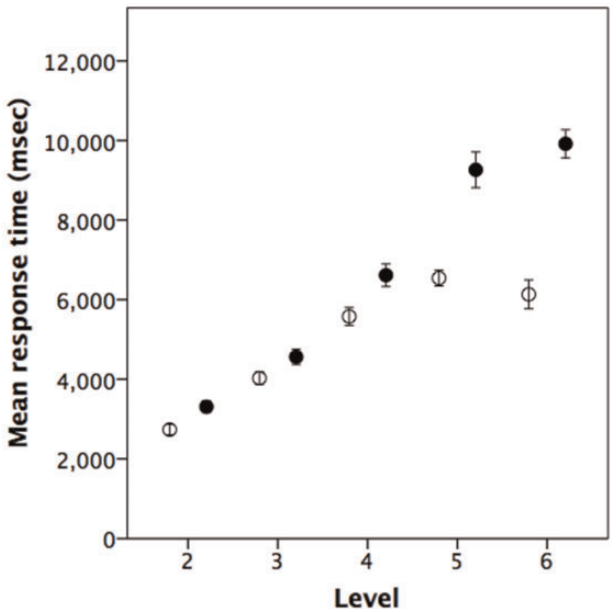


**Fig. 2.** Experiment 1: Mean reaction times of subjects when correctly answering questions at each level on mentalising (solid symbols) or factual (open symbols) recall ($N = 39$ subjects). Error bars are ± 1 SEM.

Even though those who get many questions wrong have shorter reaction times on both types of question (probably because they are guessing), reaction times vary significantly across answer categories only for mentalising questions (ANOVA with number of correct questions as independent variable: factual questions: $F_{8,186} = 1.38$, $P = 0.210$; mentalising questions: $F_{8,186} = 3.00$, $P = 0.003$), while the variances do not differ significantly across categories in either case (Levene's test for homogeneity of variances: factual questions: $F_{8,186} = 0.62$, $P = 0.759$; mentalising questions: $F_{8,186} = 1.18$, $P = 0.315$) (Figure 3). More importantly, performance on mentalising questions was always slower than performance on factual questions even when matched for the frequency of correct responses, no matter how many questions participants got wrong.

RT data are also available from the fMRI study (Experiment 2) itself. Although the constraints of the imaging design make the results difficult to interpret, we nonetheless present them here for completeness. Mean accuracy across the three question levels in the fMRI experiment was 78.7% ± 2.2 se for mentalising tasks and 79.9% ± 1.0 se for factual tasks ($F_{1,94} = 0.21$, $P = 0.650$). The mean rate of 'don't know' responses was 10.3 ± 12.8%, (10.7 ± 14.2% mentalising and 9.9 ± 11.4% memory). A 3(level) × 2 (task) ANOVA on performance accuracy revealed a significant effect of level ($F_{2,90} = 9.4$, $P < 0.001$) but not of question type (mentalising *vs* factual: $F_{1,90} = 0.2$, $P = 0.621$), with no interaction effect ($F_{2,90} = 1.3$, $P = 0.177$). Although we have reaction time data for this experiment, the experimental protocol was designed to remove effects due to difficulty (only mentalising levels 2–4 were tested, and participants prepared their response for 5–7 s while the statement was being presented prior to being cued for a response). Reaction times were slightly faster for factual questions (1588.5 ± 37.5 se ms for mentalising *vs* 1586.3 ± 25.0 se ms for factual questions). The fact that there were no main effects (level: $F_{2,90} = 0.4$, $P = 0.647$, $\eta^2 = 0.005$; question type: $F_{1,90} = 0.0$, $P = 0.988$, $\eta^2 < 0.001$) and no interaction ($F_{2,90} = 0.3$, $P = 0.707$) in these over-prepared reaction times isn't really meaningful. A fairer representation of conditions in this experiment is, perhaps, provided by the pilot experiment, since this was run with the same design but as a reaction time task in order to parameterise timings for the imaging design. Although the sample size was small ($N = 8$ only), the results (see

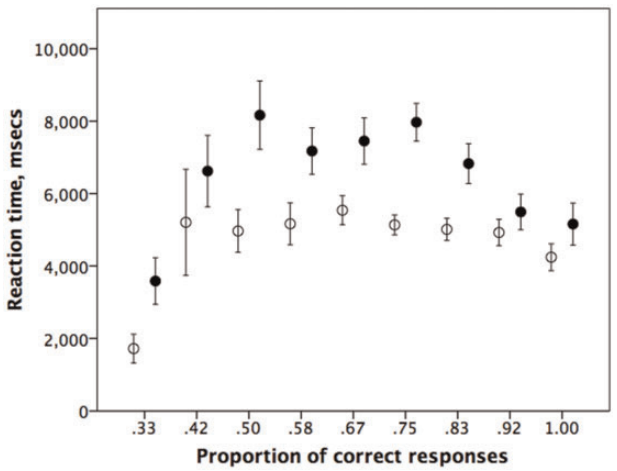


**Fig. 3.** Mean (±SE) for reaction time (in ms) as a function of the proportion of questions correctly answered at any given mentalising or factual level, for mentalising (solid symbols) *vs* factual (unfilled symbols). Data from Experiment 1.

Supplementary Data and Figure S1) are consistent with those in Figure 2.

Taking the three samples together yields a significantly consistent trend in the same direction (Fisher's meta-analysis: $\chi^2 = 25.3$, df $= 2 \times 3 = 6$, $P = 0.0003$), confirming a consistent underlying pattern.

### fMRI results

To explore the physiological implications, we carried out an fMRI experiment using the same paradigm as the behavioural experiment, except that, in order to ensure that the stimuli were well within the cognitive abilities of all participants, we included only intentionality levels 2–4 so as to be well within the natural range for adults. To determine which brain areas were involved in the mentalising task, we first pooled all intentionality levels and subtracted activations associated with the factual questions from those associated with the mentalising questions (*contrast* 1: mentalising all > factual all). This revealed robust responses in a number of regions which have been previously associated with theory of mind, including left TPJ and left dMPFC (Figure 4A). A full list of results at $P < 0.05$, whole brain corrected, is given in Table 1.

To isolate regions which were more strongly activated for more difficult mentalising tasks, we examined the parametric regressors to identify those regions in which responses were significantly modulated by intentionality level. We isolated regions which were more strongly parametrically modulated for mentalising than for factual memory (contrast 4: parametric mentalising > parametric factual). This revealed strong responses in the left insula, left posterior STS, left temporal pole, and right cerebellar hemisphere (Figure 4B; Table 1B for results at $P < 0.05$, whole brain corrected). Note that these effects are parametric functions of level, clearly demonstrating that more demanding intentionality tasks are correlated with proportionately stronger responses in these areas than is the case for memory for the facts of the story. There were no significant responses for the inverse contrast (parametric factual > parametric mentalising). For completeness, we also isolated areas that were parametrically modulated for factual memory (contrast 3) and those that were more strongly parametrically modulated for mentalising (contrast 4). Results are reported in Table 1.

In order to determine the extent to which these parametrically modulated responses overlapped with general responses to intentionality level (*contrast* 1), we plotted both sets of results on the same brain. This showed overlap in all four areas where parametrically modulated responses to intentionality had been identified, although overlap in insula involved just one voxel (Figure 4C; see Table 1B for results at $P < 0.05$, whole brain corrected) (*contrast* 1).

## Discussion

Although mentalising questions did not differ from factual questions in terms of the subject's ability to arrive at the correct answer, in both of the reaction time experiments (Experiment 1 and the pilot for Experiment 2), the mentalising questions at any given level of complexity required more cognitive processing to

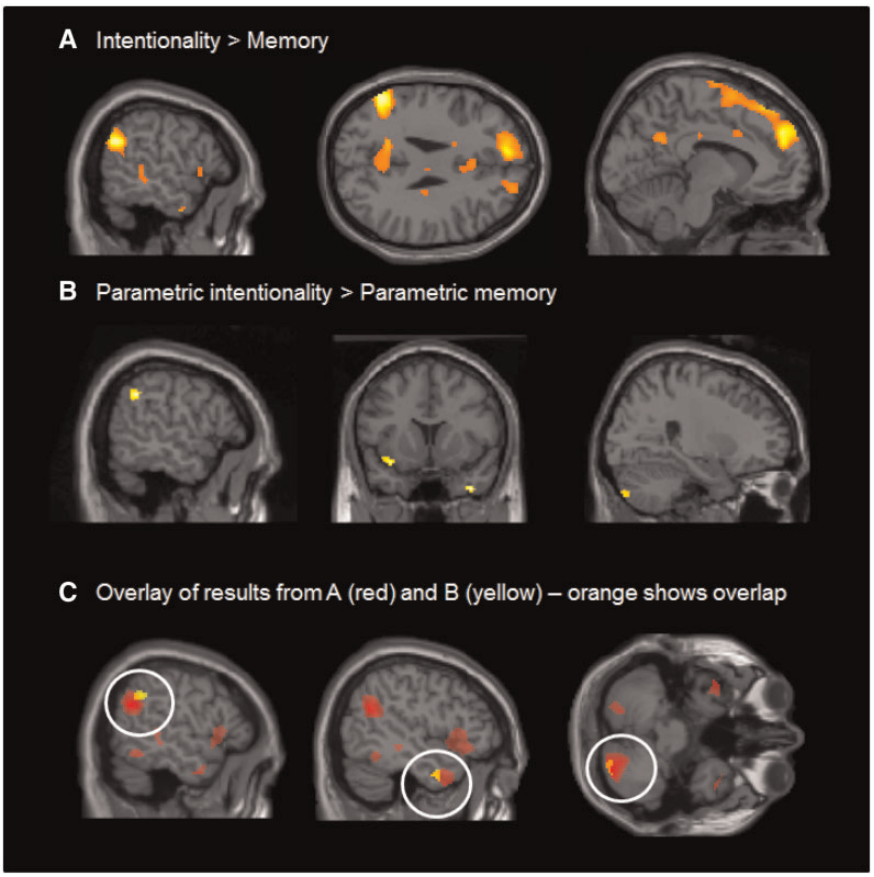


**Fig. 4.** Experiment 2: fMRI results showing (A) a broad pattern of response to the contrast intentionality > memory, (B) a more circumscribed response to the parametric modulation of difficulty levels in the intentionality *vs* factual memory tasks, and (C) the results from A (in yellow) and B (in red) plotted together. All responses shown are significant at $P < 0.05$ whole-brain corrected, as specified in the methods section.

**Table 1.** Summary of significant fMRI responses from Experiment 2 at $P < 0.001$, $k > 38$, which provides a whole brain corrected probability of $P < 0.05$

| k | equivZ | x,y,z (mm) | |
|---|---|---|---|
| A) Mentalising > Factual memory (not parametric) | | | |
| 874 | 6.9 | 28 -84 −42 | Posterior cerebellum |
| 1352 | 7.6 | −52 −58 26 | TPJ (superior temporal gyrus/supramarginal gyrus) |
| 2093 | 5.5 | −8 50 34 | Dorsomedial prefrontal cortex |
| 461 | 5.2 | 46 8 −30 | TP (middle temporal gyrus) |
| 302 | 5.1 | −46 8 −44 | TP (middle temporal gyrus) |
| 221 | 4.6 | 18 −42 76 | Postcentral gyrus |
| 606 | 4.4 | −36 0 −10 | Insula/inferior frontal gyrus |
| 120 | 4.4 | −22 −74 −42 | Posterior cerebellum |
| 488 | 4.3 | 44 −4 −10 | Insula/inferior frontal gyrus |
| 391 | 4.2 | −16 −52 28 | Precuneus |
| 149 | 4.1 | 2 −18 34 | Middle cingulate gyrus |
| 96 | 3.9 | −36 20 −26 | TP (superior temporal gyrus) |
| 46 | 3.8 | −12 16 −2 | Caudate |
| 83 | 3.7 | 0 22 26 | Anterior cingulate |
| 120 | 3.6 | −48 −32 −8 | Middle temporal gyrus |
| 41 | 3.6 | 20 8 8 | Putamen |
| B) Mentalising parametric > Factual memory parametric | | | |
| 39 | 5.7 | −34 12 16 | Insula |
| 43 | 4.1 | −46 −2 −30 | temporal pole (middle temporal gyrus) |
| 64 | 5.8 | −56 −48 34 | Temporo-parietal junction (supramarginal gyrus) |
| C) Common areas in A and B | | | |
| 38 | – | 26 −80 −44 | Posterior cerebellar lobe (declive and tuber) |
| 32 | – | −50 −4 −32 | Temporal pole (middle temporal gyrus) |
| 39 | – | −52 −52 34 | Temporo-parietal junction (supramarginal gyrus) |
| 1 | – | −40 10 −12 | Insula |
| D) Mentalising parametric | | | |
| 1990 | 5.3 | −46 −16 20 | Insula |
| 166 | 4.4 | −32 −32 −4 | Hippocampus |
| 55 | 4.2 | 66 −10 8 | Temporal pole (superior temporal gyrus) |
| 139 | 4.2 | −6 −22 80 | Postcentral gyrus/precentral gyrus |
| 46 | 4 | −26 38 −6 | Orbitofrontal cortex (middle frontal gyrus) |
| 53 | 3.9 | 50 −36 58 | Postcentral gyrus |
| 277 | 3.9 | 24 −66 −26 | Posterior cerebellum |
| C) Factual memory parametric | | | |
| 4003 | 5.4 | −22 −80 0 | Occipetal lobe |
| 318 | 4.9 | −22 36 −8 | Orbitofrontal cortex (middle frontal gyrus) |
| 259 | 4.4 | 40 −38 66 | Postcentral gyrus |
| 73 | 4.4 | 12 −66 −50 | Cerebellum |
| 187 | 4.3 | 52 −66 −10 | Middle occipetal gyrus |
| 360 | 4.3 | 66 −12 10 | Supramarginal gyrus |
| 250 | 4.2 | 46 −66 −28 | Cerebellar hemisphere |
| 74 | 4.1 | −44 −72 −28 | Cerebellar hemisphere |
| 73 | 3.9 | 34 2 16 | Insula |
| 52 | 3.7 | 28 −66 4 | Middle occipetal gyrus |
| 63 | 3.5 | 2 −56 −26 | Cerebellar vermis |
| 46 | 3.5 | 18 −68 −26 | Mid cerebellum o |
| 44 | 3.5 | −34 −64 −24 | Cerebellar hemisphere |

achieve this. More importantly, there was a parametric effect of question level, with mentalising questions becoming progressively more taxing than factual memory questions as their order increased. This was not true of reaction times in the fMRI experiment (Experiment 2) itself, as participants were allowed to overprepare and it is questionable as to what these results actually tell us; nonetheless, overall across the three sets of experimental data, the results were in the same direction. Note that, in Experiment 1, reaction times for the factual recall task seemed to become asymptotic after level 4, but continued to rise for the mentalising tasks (Figure 2). (We cannot tell if this also happened in the two datasets from Experiment 2 as only levels 2–4 were considered in this case.) This would seem to reinforce the claim that mentalising recall tasks become progressively more demanding than factual recall tasks that lacked a mentalising component: factual recall tasks do not necessarily continue to

increase in difficulty, but mentalising tasks do. These results were supported by the fMRI experiment, which suggested that, despite the reaction time results, mentalising tasks recruit more neural response than simple factual recall tasks, and do so disproportionately as intentionality level increases compared to non-mentalising tasks matched for factual complexity.

In addition, our results confirm (i) that mentalising tasks are cognitively more demanding than factual tasks and, more importantly, (ii) that there is a significant parametric effect in the brain regions involved as a function of the intentionality level at which subjects are required to work (higher order tasks require the recruitment of disproportionately more neural effort compared to equivalent non-mentalising tasks).

Although fMRI studies have consistently implicated a network of regions in the temporal lobe and prefrontal cortex in theory of mind reasoning (Carrington and Bailey 2009; van Overwalle, 2009), the claim that mentalising itself is especially cognitively demanding has never actually been tested, despite the fact that it is a core assumption of both the social brain hypothesis (Byrne and Whiten, 1988; Dunbar, 1998) and the social (or communicative) complexity hypothesis (Freeberg *et al.*, 2012). The social brain hypothesis argues that the kinds of social decisions necessitated by the more complex social arrangements found in species like anthropoid primates are more cognitively demanding, and that this is reflected in the need for larger brains (Byrne and Whiten, 1988), and hence a correlation between frontal lobe volume, in particular, and social group size in primates (Dunbar and Shultz, 2007).

Although the precise relationship between cognitive processing demand and brain volume (or neural density) remains undetermined, an agent based model has suggested that, as implied by the social complexity hypothesis, the more complex social decisions required to support larger social groups are 'cognitively' more demanding (as indexed by CPU processing time) (Dávid-Barrett and Dunbar, 2013). In this respect, the results reported in this paper suggest one way in which the social brain hypothesis might have been instantiated: the greater cognitive processing required for more complex decision making is achieved by increasing the amount of neural tissue available for this. Our results do not, of course, provide direct evidence for this, but they are at least consistent with such a conclusion. As an hypothesis, we might suggest that the additional cognitive demand arises from the need to model other individuals' mental states in a virtual world rather than simply relying on direct physical cues (or simple association learning).

Most of the regions identified as disproportionately implicated in mentalising by our imaging study are those already known to be part of the 'theory of mind network'. Our results indicate that this is a quantitative effect rather than just a qualitative one: activity in these regions increases disproportionately with mentalising level. In addition, however, our analyses also suggest that other brain regions not normally associated with mentalising may also play a role. The cerebellum was one such identified by the contrast analysis (Figure 4B). Although not typically thought of as critical for social cognition, the cerebellum has been linked to both autism (Baron-Cohen *et al.*, 1999; Ito, 2008) and basic theory of mind processing (Brothers and Ring, 1992). The cerebellar role in cognitive tasks of this type is not well understood, but the cerebellum is widely thought to play a role in managing integration across different cognitive processes (Kolb and Wishaw, 1996; Ramnani, 2006; Ito, 2008; Wiestler *et al.*, 2011; Koziol *et al.*, 2012). This being so, it might prove to be especially important when managing several different mind states simultaneously, especially when it is necessary to keep these differentiated as in mentalising tasks of the kind considered here. Keeping track of one other mindstate (in addition to one's own) may not be so challenging, but managing three others may be and so may demand cerebellar input.

It is important to note that our baseline measure is the number of factual propositions in each vignette, not the embeddedness of these factual elements. Embeddedness is, by definition, a property of mentalising, and the present design does not dissociate these two components. For present purposes, however, we are less concerned with the difference between mentalising and embeddedness than with the difference between mentalising (in effect, the number of mentalised facts in a story, embedded or otherwise across individual minds) and simple factual memory (the number of non-mentalised facts in the same story). Indeed, grammatical embeddedness may be one way in which high order mentalising is scaffolded. Nonetheless, other experiments have suggested that mentalising may be more limiting of individuals' abilities to process complex sentences than grammatical embeddeness per se (Oesch and Dunbar, 2017).

An alternative possible source of confound is that mentalising and factual questions might differ in the ease with which a participant can identify the proposition that makes a sentence false. In all our questions, each false statement had only one false element (clause), and all the rest were true. It could therefore be that a false mentalising question requires one to read right to the end of the sentence in order to know that it was false, whereas a false factual question can be identified as soon as one encounters the false proposition. If this were true, then the variance on factual questions should be greater than that on mentalising questions because the false statement could be in any position from first to last clause in the question, whereas mentalising questions would require one to read to the end. Since the variances are in fact very small and do not differ between the two kinds of questions (Figure 2; Supplementary Figure S1), this cannot explain our results. An alternative possibility might be that false mentalising questions can be identified as soon as one comes across the false proposition, whereas one would have to read the whole of a factual question to be sure that no facts were incorrect. However, if this was true, then we would expect exactly the opposite results to those we obtained (i.e. mentalising questions would be processed faster).

There are, perhaps, several reasons why these potential sources of confound are unlikely to explain the results. One is that the error variance does not vary systematically with question level, or between question types (Figure 2). Secondly, half the questions were true questions and half false, but the position effect can only apply to false questions; subjects would have to read to the end for all true questions of both types to make sure that the false proposition was not in the last clause. Hence, the bias would have to massively increase the variance in response time on false questions in order to compensate for the lack of effect on the true questions. Finally, and perhaps, more importantly, a plot of reaction time as a function of performance does not suggest that questions become differentially faster to process the more accurately they are answered (Figure 3). This suggests that the differences we observe are most likely due to the added cognitive demands of mentalising.

The functional imaging results demonstrate that, as subjects need to simultaneously judge more mental states and the relationships between the individuals involved, they draw on progressively greater neural resources, sometimes involving a larger number of brain regions, especially in those brain regions within the frontal and temporal lobes that are known to be associated with theory of mind. This need to recruit additional neural power might help to explain why neocortex (and

particularly frontal lobe) volume correlates both with social group size across primates (Dunbar, 1998; Dunbar and Shultz, 2007) and with personal social network size in humans (Lewis *et al.*, 2011; Kanai *et al.*, 2012; Powell *et al.*, 2012, 2014) and macaques (Sallet *et al.*, 2011). However, the question of whether cognitive demand is reflected in neural volume remains to be determined.

In summary, these findings suggest that social cognitive processing (mentalising) is unusually computationally demanding, thereby confirming of a central assumption of the social brain hypothesis. This may have particular relevance to the kinds of cognition that are peculiar to anthropoid primates (Passingham and Wise 2011) and which may, in quantitative terms at least, be especially unique for humans.

## Acknowledgements

This study was made possible by a University of Liverpool VIP award. We thank Rachel Browne for help in pilot data collection, Steve Platek for design advice and Chris Frith for comments on an earlier version of the manuscript. RD's research is supported by a European Research Council Advanced Investigator grant.

## Supplementary data

Supplementary data are available at *SCAN* online.

*Conflict of interest*. None declared.

## References

Baillargeon, R., Scott, R.M., He, Z. (2010). False-belief understanding in infants. *Trends in Cognitive Sciences*, **14**, 110–8.

Byrne, R.W., Whiten, A. (1988) *The Machiavellian Intelligence Hypothesis*. Oxford: Oxford University Press.

Baron-Cohen, S., Ring, H.A., Wheelwright, S., *et al.* (1999). Social intelligence in the normal and autistic brain: an fMRI study. *European Journal of Neuroscience*, **11**, 1891–8.

Barrett, L., Henzi, S.P., Dunbar, R.I.M. (2003). Primate cognition: from 'what now?' to 'what if?'. *Trends in Cognitive Sciences*, **7**, 494–7.

Brothers, L., Ring, B. (1992). A neuroethological framework for the representation of minds. *Journal of Cognitive Neuroscience*, **4**, 107–18.

Carrington, S.J., Bailey, A.J. (2009). Are there Theory of Mind regions in the brain? A review of the neuroimaging literature. *Human Brain Mapping*, **30**, 2313–35.

Corallo, G., Sackur, J., Dehaene, S., Sigman, M. (2008). Limits on introspection: distorted subjective time during the dual-task bottleneck. *Psychological Science*, **19**, 1110–7.

Dávid-Barrett, T., Dunbar, R.I.M. (2013). Processing power limits social group size: computational evidence for the cognitive costs of sociality. *Proceedings of the Royal Society of London*, 20131151.

Deichmann, R., Gottfried, J.A., Hutton, C., Turner, R. (2003). Optimized EPI for fMRI studies of the orbitofrontal cortex. *NeuroImage*, **19**, 430–41.

Dennett, D. (1983). Intentional systems in cognitive ethology: the "Panglossian paradigm" defended. *Behavioral and Brain Sciences*, **6**, 343–90.

Dunbar, R.I.M. (1998). The social brain hypothesis. *Evolutionary Anthropology*, **6**, 178–90.

Dunbar, R.I.M. (2010) Evolutionary basis of the social brain. In: Decety, J., Cacioppo, J., editors. *Oxford Handbook of Social Neuroscience*. Oxford: Oxford University Press, pp. 28–38.

Dunbar, R.I.M., Shultz, S. (2007). Understanding primate brain evolution. *Philosophical Transactions of the Royal Society of London*, **362B**, 649–58.

Freeberg, T., Dunbar, R.I.M., Ord, T. (2012). Social complexity as a proximate and ultimate factor in communicative complexity. *Philosophical Transactions of the Royal Society of London*, **367B**, 1785–801.

Frith, U., Frith, C.D. (2003). Development of a neurophysiology of mentalising. *Philosophical Transactions of the Royal Society of London*, **358B**, 459–73.

Fukui, H., Murai, T., Shinozaki, J., *et al.* (2006). The neural basis of social tactics: an fMRI study. *NeuroImage*, **32**, 913–20.

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

Gallagher, H.L., Jack, A.I., Roepstorff, A., Frith, C.D. (2002). Imaging the intentional stance in a competitive game. *NeuroImage*, **16**, 814–21.

Henzi, P., de Sousa Pereira, L., Hawker-Bond, D., Stiller, J., Dunbar, R.I.M., Barrett, L. (2007). Look who's talking: developmental trends in the size of conversational cliques. *Evolution and Human Behavior*, **28**, 66–74.

Ito, M. (2008). Control of mental activities by internal models in the cerebellum. *Nature Reviews Neuroscience*, **9**, 304–13.

Kanai, R., Bahrami, B., Roylance, R., Rees, G. (2012). Online social network size is reflected in human brain structure. *Proceedings of the Royal Society of London*, **279B**, 1327–34.

Kinderman, P., Dunbar, R.I.M., Bentall, R.P. (1998). Theory-of-mind deficits and causal attributions. *British Journal of Psychology*, **89**, 191–204.

Kobayashi, C., Glover, G.H., Temple, E. (2006). Cultural and linguistic influence on the neural basis of 'Theory of Mind': an fMRI study with Japanese bilinguals. *Brain and Language*, **98**, 210–20.

Kolb, B., Wishaw, I.Q. (1996) *Fundamentals of Human Neuropsychology*. San Francisco: W.H. Freeman.

Kovács, ÁM., Téglás, E., Endress, A.D. (2010). The social sense: susceptibility to others' beliefs in human infants and adults. *Science*, **330**, 1830–4.

Koziol, L.F., Budding, D.E., Chidekel, D. (2012). From movement to thought: executive function, embodied cognition, and the cerebellum. *Cerebellum*, **22**, 505–25.

Lewis, P., Rezaie, R., Browne, R., Roberts, N., Dunbar, R.I.M. (2011). Ventromedial prefrontal volume predicts understanding of others and social network size. *NeuroImage*, **57**, 1624–9.

Lin, S., Keysar, B., Epley, N. (2010). Reflexively mindblind: using theory of mind to interpret behavior requires effortful attention. *Journal of Experimental Social Psychology*, **46**, 551–6.

Meyer, M.L., Spunt, R.P., Berkman, E.T., Taylor, S.E., Lieberman, M.D. (2012). Evidence for social working memory from a parametric functional MRI study. *PNAS*, **109**, 1883–8.

Oesch, N., Dunbar, R.I.M. (2017). The emergence of recursion in human language: mentalising predicts recursive syntax task performance. *Journal of Neurolinguistics*, doi: http://dx.doi.org/10.1016/j.jneuroling.2016.09.008.

Perner, J. (1991) *Understanding the Representational Mind*. Cambridge (MA): MIT Press.

Powell, J., Lewis, P., Dunbar, R.I.M., García-Fiñana, M., Roberts, N. (2010). Orbital prefrontal cortex volume correlates with social cognitive competence. *Neuropsychologia*, **48**, 3554–62.

Passingham, R.E., Wise, S.P. (2011) *The Neurobiology of the Prefrontal Cortex*. Oxford: Oxford University Press.

Powell, J., Lewis, P.A., Roberts, N., García-Fiñana, M., Dunbar, R.I.M. (2012). Orbital prefrontal cortex volume predicts social network size: an imaging study of individual differences in humans. *Proceedings of the Royal Society of London*, **279**, 2157–62.

Powell, J., Kemp, G., Dunbar, R.I.M., Roberts, N., Sluming, V., García-Fiñana, M. (2014). Different association between intentionality competence and prefrontal volume in left- and right-handers. *Cortex*, **54**, 63–76.

Premack, D., Woodruff, G. (1978). Does the chimpanzee have theory of mind. *Journal of Behavioral and Brain Science*, **1**, 515–26.

Ramnani, N. (2006). The primate cortico-cerebellar system: anatomy and function. *Nature Reviews Neuroscience*, **7**, 511–22.

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

Sallet, J., Mars, R., Noonan, M., *et al.* (2011). Social network size affects neural circuits in macaques. *Science*, **334**, 697–700.

Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology*, **16**, 235–9.

Saxe, R., Powell, L.J. (2006). It's the thought that counts: specific brain regions for one component of theory of mind. *Psychological Science*, **17**, 692–9.

Schneider, D., Lam, R., Bayliss, A.P., Dux, P.E. (2012). Cognitive load disrupts implicit theory-of-mind processing. *Psychological Science*, **23**, 842–7.

Stiller, J., Dunbar, R.M. (2007). Perspective taking and memory capacity predict social network size. *Social Networks*, **29**, 93–104.

Tomasello, M., Call, J. (1998) *Primate Cognition*. New York: Academic Press.

van Overwalle, F. (2009). Social cognition and the brain: a meta-analysis. *Human Brain Mapping*, **30**, 829–58.

Vollm, B.A., Taylor, A.N.W., Richardson, P., *et al.* (2006). Neuronal correlates of theory of mind and empathy: a functional magnetic resonance imaging study in a nonverbal task. *NeuroImage*, **29**, 90–8.

Wiestler, T., McGonigle, D.J., Diedrichsen, J. (2011). Integration of sensory and motor representations of single fingers in the human cerebellum. *Journal of Neurophysiology*, **105**, 3042–53.