OXFORD

Social Cognitive and Affective Neuroscience, 2016, 1687–1697

doi: 10.1093/scan/nsw098
Advance Access Publication Date: 21 July 2016
Original article

Specificity, reliability and sensitivity of social brain responses during spontaneous mentalizing

Carolin Moessnang,[1] Axel Schäfer,[1] Edda Bilek,[1] Paul Roux,[2,3,4,5] Kristina Otto,[1] Sarah Baumeister,[6] Sarah Hohmann,[6] Luise Poustka,[6,7] Daniel Brandeis,[6,8,9,10] Tobias Banaschewski,[6] Andreas Meyer-Lindenberg,[1] and Heike Tost[1]

[1]Department of Psychiatry and Psychotherapy, Systems Neuroscience in Psychiatry, Central Institute of Mental Health, Medical Faculty Mannheim/Heidelberg University, Mannheim, Germany, [2]Laboratoire de Sciences Cognitives et Psycholinguistique, UMR 8554, CNRS-ENS-EHESS, Institut d'Étude de la Cognition, Ecole Normale Supérieure, Paris, France, [3]Service Universitaire de Psychiatrie d'adultes, Centre Hospitalier de Versailles, Le Chesnay, France, [4]Laboratoire HandiRESP EA4047, Université Versailles Saint Quentin En Yvelines, Versailles, France, [5]Fondation FondaMental, Créteil, France, [6]Department of Child and Adolescent Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim/Heidelberg University, Mannheim, Germany, [7]Department of Child and Adolescent Psychiatry, Medical University of Vienna, Vienna, Austria, [8]Department of Child and Adolescent Psychiatry and Psychotherapy, University Hospital of Psychiatry Zurich, Zurich, Switzerland, [9]Center for Integrative Human Physiology, University of Zurich, Zurich, Switzerland and [10]Neuroscience Center Zurich, ETH and University of Zurich, Zurich, Switzerland

Correspondence should be addressed to Carolin Moessnang, PhD, Systems Neuroscience in Psychiatry, Department of Psychiatry and Psychotherapy, Central Institute of Mental Health, Medical Faculty Mannheim/Heidelberg University, J5, 68159 Mannheim, Germany. Email: carolin.moessnang@zi-mannheim.de

Abstract

The debilitating effects of social dysfunction in many psychiatric disorders prompt the need for systems-level biomarkers of social abilities that can be applied in clinical populations and longitudinal studies. A promising neuroimaging approach is the animated shapes paradigm based on so-called Frith-Happé animations (FHAs) which trigger spontaneous mentalizing with minimal cognitive demands. Here, we presented FHAs during functional magnetic resonance imaging to 46 subjects and examined the specificity and sensitivity of the elicited social brain responses. Test–retest reliability was additionally assessed in 28 subjects within a two-week interval. Specific responses to spontaneous mentalizing were observed in key areas of the social brain with high sensitivity and independently from the variant low-level kinematics of the FHAs. Mentalizing-specific responses were well replicable on the group level, suggesting good-to-excellent cross-sectional reliability [intraclass correlation coefficients (ICCs): 0.40–0.99; dice overlap at $P_{uncorr}<0.001$: 0.26–1.0]. Longitudinal reliability on the single-subject level was more heterogeneous (ICCs of 0.40–0.79; dice overlap at $P_{uncorr}<0.001$: 0.05–0.43). Posterior temporal sulcus activation was most reliable, including a robust differentiation between subjects across sessions (72% of voxels with ICC>0.40). These findings encourage the use of FHAs in neuroimaging research across developmental stages and psychiatric conditions, including the identification of biomarkers and pharmacological interventions.

Key words: spontaneous mentalizing; fMRI; animated shapes; reliability; biomarker

Received: 16 March 2016; **Revised:** 28 May 2016; **Accepted:** 11 July 2016

Introduction

Human mentalizing, also referred to as theory of mind (ToM), includes the recognition of intentions and emotions of a social partner, an ability which has convincingly been related to the interaction of a circumscribed set of neural regions commonly assigned to the 'social brain' (e.g. posterior superior temporal sulcus [pSTS], dorsomedial pre-frontal cortex [dmPFC]; Adolphs, 2009). The in-depth characterization of this network with functional magnetic resonance imaging (fMRI) is a major goal of contemporary social and clinical neuroscience. This is reflected by big-data consortia aiming at the identification of the neural mechanisms of social cognition and associated behavioral deficits such as those observed in autism spectrum disorders (ASD; e.g. EU-AIMS consortium; Murphy and Spooren, 2012). A critical prerequisite for suitable neuroimaging biomarkers for developmental and pharmacological research is the assessment of the specificity, reliability and sensitivity of neural responses elicited by mentalizing fMRI paradigms (Loth *et al.*, 2015). However, despite the broad recognition of these requirements (Bennett and Miller, 2010) and the wealth of published fMRI tasks on mentalizing functions (Schurz *et al.*, 2014), no such data is available to date.

Here, we developed an approach to provide systems-level biomarkers of this kind using the so-called Frith-Happé animations (FHAs), a promising candidate tool for the assessment of social brain activations across a broad range of developmental ages and disorder severities. The FHAs are a set of validated video clips depicting simple geometric shapes which interact at different levels of social significance and prompt spontaneous mentalizing functions (Abell *et al.*, 2000). This form of mentalizing occurs automatically, is observable already at an early age, and forms the basis of social functioning in everyday life (Mar and Macrae, 2007; Apperly and Butterfill, 2009). The relevance of spontaneous mentalizing abilities is exemplified by observations that ASD patients can learn to pass other mentalizing tasks based on explicit reasoning, but have difficulties in spontaneous, i.e. more implicitly driven ToM (Frith, 2004; Senju *et al.*, 2009; Schneider *et al.*, 2013). While the neural signals underlying spontaneous mentalizing clearly map to the social brain (Castelli *et al.*, 2000), questions regarding the specificity and quality criteria of FHA-induced brain responses remain to be addressed.

The current study therefore investigates the specificity, reliability and sensitivity of neural responses to the FHAs. Based on previous evidence (Castelli *et al.*, 2000), we hypothesized that FHA-induced activation of the social brain can be robustly reproduced in a well-powered sample. In order to address the specificity of brain activation in the light of potential confounds, we hypothesized that neural responses to the FHAs in higher order social brain areas are specific to mentalizing processes (as compared to sub-ordinate processes related to animacy and agency perception) and unrelated to the low-level kinematic properties of the animated video clips (Roux *et al.*, 2013). Using a test–retest approach, the reliability of neural responses was assessed in order to inform the use of FHAs in different experimental designs: While cross-sectional reliability is critical for designs involving group comparisons, such as case-control studies (e.g. in imaging genetics or pharmaco-fMRI studies with parallels group designs), longitudinal reliability is required for repeated measures analyses in developmental and cross-over pharmacological designs (Bennett and Miller, 2010). A third type of reliability refers to the aspect of between-subject reliability, which reflects the replicability of between-subject differences and is particularly desirable in endophenotype research (Raemaekers *et al.*, 2007). Since the 'bottom up'-like nature of spontaneous mentalizing is suggestive of a fairly stable response of the underlying neural circuitry, we expected fair to good (Shrout and Fleiss, 1979) reliability of FHA-induced brain activations. In addition, statistical sensitivity was assessed as a complementary measure and assumed to be high as a result of the block design structure of the task.

Materials and methods

Experimental procedure

Participants. A total of 46 healthy volunteers (mean age: 24.7 ± 5.3 years, 21 females) participated in the study, a subsample of which ($n = 28$, mean age: 22.9 ± 2.8 years, 14 females) performed fMRI scanning twice in order to assess test–retest reliability estimates. Exclusion criteria included a lifetime history of neurological or psychiatric disorder, current intake of psychoactive substances, significant general medical problems including liver, cardiac, or renal dysfunctions, a history of head trauma, and pregnancy. All individuals provided written informed consent for a study protocol that was approved by the institutional review board of the Medical Faculty Mannheim.

Paradigm. We integrated the FHAs (Abell *et al., 2000*) in a block-designed fMRI paradigm with a pseudorandomized order of task conditions [random (R), goal-directed (GD), ToM]. All video clips featured a big and a little triangle moving about the screen. In the ToM condition, the triangles' movement patterns suggested complex intentional interactions challenging the observer's mentalizing abilities (e.g. one triangle deceiving the other). In the higher order control condition, the triangles interacted purposefully, thereby conveying the perception of agency, but did not challenge mentalizing functions (GD, e.g. one triangle imitating the other). In the second, more basic control condition (not analyzed further here), the triangles moved randomly without interaction (R), thereby only allowing for the perception of animacy. Following the procedures validated by (White *et al., 2011*), we asked subjects to indicate the subjective social significance level of the depicted interactions (i.e. ToM, GD or R) after each video clip presentation (multiple-choice questions for categorization; MCQ-cat). After the social significance ratings of ToM video clips, we additionally asked the subjects to rate the perceived emotional state (i.e. positive, neutral or negative emotional valence) of each triangle at the end of the video clip presentation (multiple-choice question for the triangle's feeling; MCQ-feeling). This was done to assess the subjective emotional significance of the scenes and to provide an indicator of whether the expected social cognitive and emotional concept of the displayed interactions had indeed been understood. Further details on the stimuli, trial structure and ratings of the fMRI paradigm are provided in *Figure* 1. Subjects were thoroughly instructed and trained prior to fMRI scanning using three established practice video clips.

MRI data acquisition. Functional MRI was performed on a 3 T Siemens Trio Scanner (Siemens, Erlangen, Germany) equipped with a 12-channel head coil. Functional data was collected using an echo-planar imaging (EPI) sequence with the following parameters (TE: 30 ms, TR: 2 s, α: 80°, matrix: 64 × 64, FOV: 192 × 192 mm, in-plane resolution: 3 × 3 mm, slice thickness: 4 mm, gap: 1 mm, 28 axial slices, 331 volumes).

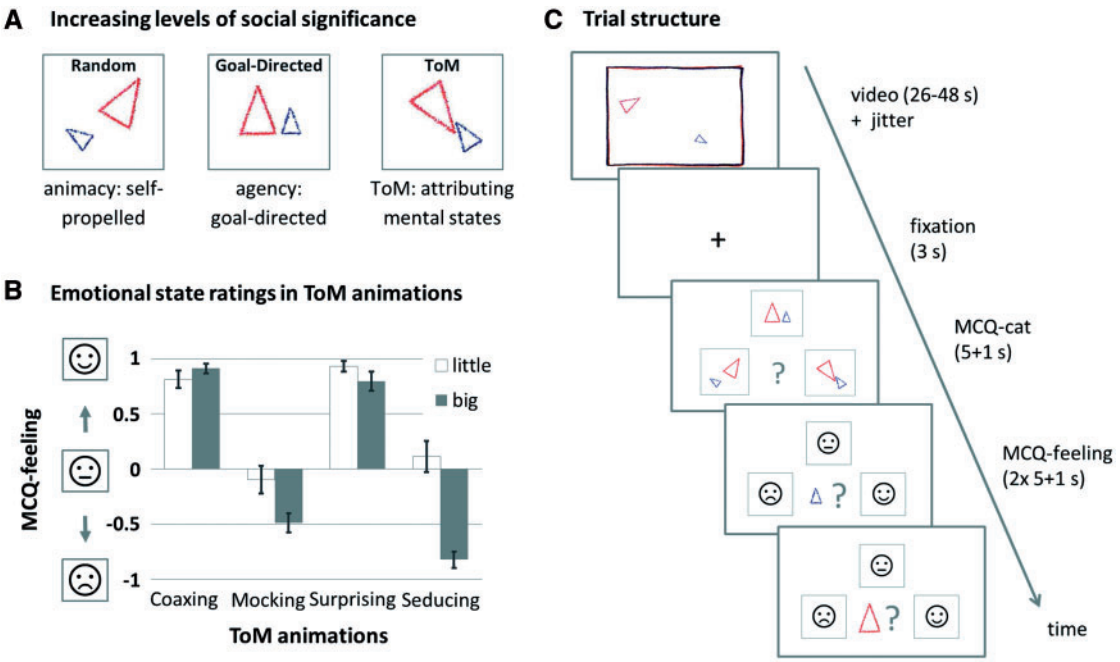

Fig. 1. Overview over stimuli and experimental design. (A) Stimuli consisted of three types of animated video clips with increasing levels of social significance, represented by three simplified icons during the subsequent categorization (MCQ-cat: 'Which category did the previously presented animation belong to?'). Example video clips are accessible at https://sites.google.com/site/utafrith/research. (B) ToM animations were additionally rated according to perceived emotionality (MCQ-feeling: 'How did the little/big triangle feel at the end of the animation?'), which probed for the acquired concept of the displayed emotional states, and thus served as a rough estimation of the subject's understanding of the cover story: 'coaxing' and 'surprising' require both triangles to be rated similarly positively, while 'mocking' and 'seducing' require the little triangle to be rated more positively than the big triangle. Emotional states were represented by schematic faces with an unhappy, neutral and happy expression, respectively. Bar graphs depict average rating scores (± SE) for each ToM animation, with '+1' referring to 'happy' and '−1' referring to 'unhappy.' Panel (C) depicts the temporal structure of a ToM trial. The presentation of the video clips was preceded by a jitter with variable duration ($M = 995.67$ ms, s. d. = 418.3). Responses were given with the right thumb, using the left, upper and right key of an MRI compatible button box (Current Designs, PA, USA). As soon as responses were given during MCQ ratings, the chosen icon was framed in red for the duration of one additional second, followed by a blank screen for the remainder of the respective MCQ phase. No feedback on response accuracy was given.

Test–retest procedures. Subjects participating in the test–retest study performed all FHAs task procedures twice, including the instructions and practice sessions (mean time interval between the first [T1] and second [T2] session: 15.8 ± 3.5 days). Basic variables such as time of day, hours of sleep, cigarettes smoked and caffeine intake were matched as closely as possible between T1 and T2 in order to control for potential physiological confounds (all P values > 0.05; Supplementary Table S1). In addition, scanner quality assurance (QA) was performed according to an established QA protocol (Plichta *et al.*, 2012). Metrics pertaining to mean signal intensity, spatial and temporal signal-to-noise ratio, percent signal fluctuations and percent signal drift were acquired on every measurement day (n = 37), using the identical EPI sequence parameters as outlined above for a total of 150 volumes. Due to technical issues, QC data was not available for four of the acquired measurements. All QA metrics were stable within the time range of the study (December 2012 to May 2013; Supplementary Table S2).

Spatial definition of social brain regions. A priori regions of interest (ROI) were defined based on meta-analytical data highlighting their well-established role in human mentalizing. To this end, ROI masks were derived from a recent influential publication on nonstory-based (i.e. non-verbal) ToM studies (Mar, 2011). The masks delineated cortical regions including the dorsomedial prefrontal cortex (dmPFC, bilateral mask size: 290 voxels), the posterior aspects of the superior temporal sulcus region (pSTS, left: 189 voxels, right: 227 voxels), precuneus (bilateral: 201 voxels), anterior middle temporal gyrus (aMTG; left: 75 voxels, right: 81 voxels), and inferior frontal gyrus (IFG, left: 84 voxels, right 104 voxels). Masks for the temporal poles (TP; left: 254, right: 391 voxels) were based on the Anatomical Automatic Labeling Atlas (Tzourio-Mazoyer *et al.*, 2002). See Supplementary Figure S1 for an illustration of the examined ROIs.

Data analysis

Functional correlates of spontaneous mentalizing. Image preprocessing followed standard processing routines in SPM8 (http://www.fil.ion.ucl.ac.uk/spm/). Briefly, data was realigned to the first image, slice time corrected, spatially normalized into standard stereotactic space defined by the Montreal Neurological Institute (MNI) template, resampled to 3 mm isotropic voxels, and smoothed with an 8 mm full-width at half-maximum Gaussian Kernel. Individual general linear models (GLMs) included condition-wise regressors, which were calculated by convolving the modelled box-car functions of video clip presentations with the standard canonical hemodynamic response function (HRF) implemented in SPM. Realignment parameters were included as covariates of no interest at the first level. During model estimation, the data was high-pass filtered with a cutoff of 256 s, and an autoregressive model of the first order was applied. In order to separate the functional correlates of mentalizing from those of lower-level processes related to animacy and agency perception, the mentalizing condition (ToM) was contrasted to the higher-level baseline condition (GD). Resulting individual contrast images were subjected to one-sample t-tests for group-level inference. Activations were

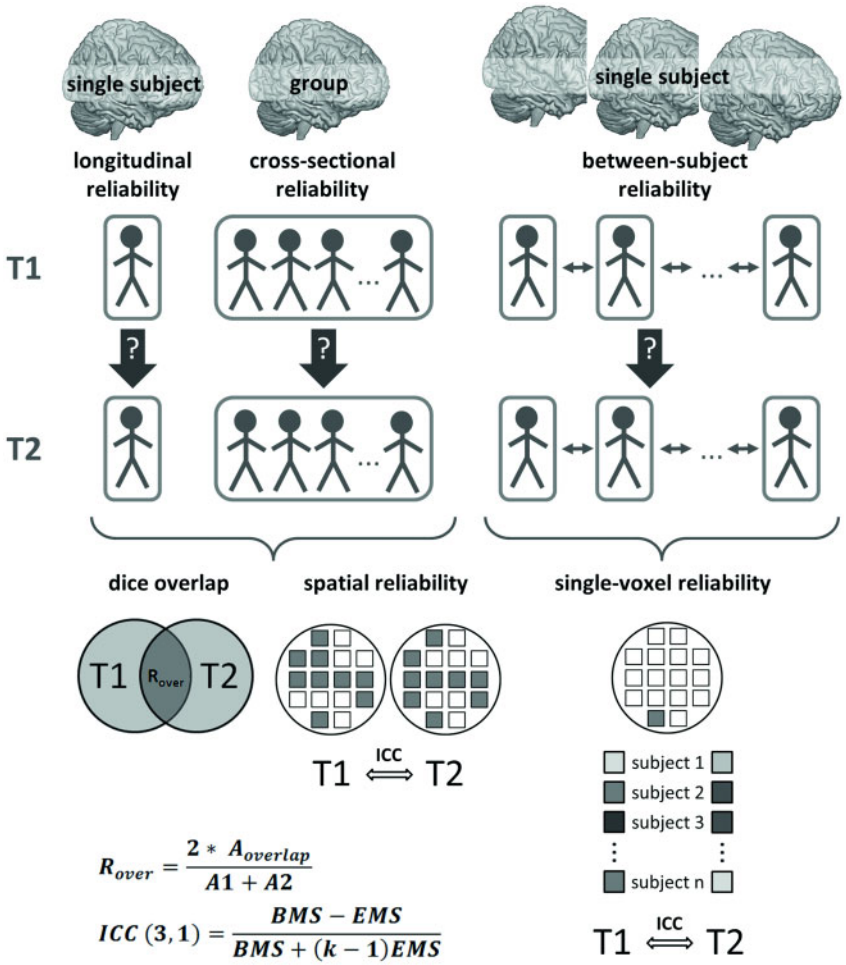

Fig. 2. Schematic overview over the reliability analysis strategy, with A1/A2: number of supra-threshold voxels at T1/T2, $A_{overlap}$: number of voxels with supra-threshold activation at both T1 and T2, BMS: between-subjects mean square, EMS: error mean square, *k*: number of repeated sessions.

reported at a significance level of $P < 0.05$, peak-level family-wise error corrected for multiple comparisons across the whole brain.

Test–retest reliability. Test–retest reliability was determined for the high-level contrast ToM > GD using the following complementary metrics (described in detail below and illustrated in *Figure 2*):

1. Dice overlap of thresholded t-maps (Rombouts *et al.*, 1998; Smith *et al.*, 2005),
2. Spatial reliability, defined as the replicability of the spatial signal intensity distribution derived from contrast maps (Raemaekers *et al.*, 2007), and
3. Single-voxel reliability, defined as the replicability of single-voxel signal intensities across subjects and time points (Caceres *et al.*, 2009).

Cross-sectional and longitudinal reliabilities resulted from the application of the metrics 'dice overlap' and 'spatial reliability' to group level and single-subject level activation maps, respectively. Between-subject reliability was assessed using the metric 'single-voxel reliability' in order to determine how well the rank order of subjects is preserved across sessions. All three metrics were applied to each ROI and to the whole-brain task network (defined by voxels surviving $P < 0.001$ at the group level at T1, uncorrected for multiple comparisons across the whole brain; Figure 2D), and calculations were performed as follows:

The intra-class correlation coefficient (ICC) with a consistency criterion (ICC(3,1); Shrout and Fleiss, 1979) was used to assess single-voxel and spatial reliability. Following Fleiss (1986), we denote ICC values < 0.40 as poor, 0.40–0.75 as fair to good and > 0.75 as excellent. Single-voxel reliability was calculated as ICC on signal intensity across subjects and time points for each voxel, and summary values were calculated as percent voxels with ICC > 0.4 for each ROI. This measure therefore quantifies the proportion of voxels which meet the reliability criterion for each ROI. Spatial reliability was based on contrast maps and was calculated as ICC on signal intensity across voxels and time points. Finally, dice overlap was based on thresholded *t*-maps calculated as $R_{OVERLAP} = 2*A_{OVERLAP}/(A1 + A2)$, which ranges from 0 (no overlap) to 1 (total overlap). The variables *A1*, *A2* and $A_{OVERLAP}$ represent the quantity of supra-threshold voxels at *T1*, *T2*, and for both sessions, respectively. The overlap was evaluated for two thresholds, $P < 0.001$ and $P < 0.005$, uncorrected for multiple comparisons across the whole brain, in order to additionally assess the dependence of the overlap measure on the applied threshold. A lack of supra-threshold activation was penalized with assigning a value of zero. In addition, following the assumption that single-subject level reliability should be reflected in higher within- than between-subject overlap (Gorgolewski *et al.*, 2013), session-wise overlap measures were

computed between each subject and all other subjects and statistically compared to within-subject overlap measures. To this end, the within-subject overlap of each individual was compared to a randomly chosen overlap measure involving the same individual (i.e. between- and within-subject overlap), and an average difference score was calculated. This comparison was repeated 10 000 times. An empirical *P* value was calculated as the number of occurrences with smaller within-subject overlap (i.e. difference score < 0), divided by the number of repetitions. *P* values < 0.05 therefore reflect significantly higher within- than between-subject overlap for the respective ROI.

Power calculations. In order to assess task sensitivity, i.e. the ability to detect an effect of task in the fMRI signal, power calculations were performed for each ROI separately as well as for all ROIs combined using the FMRIpower toolbox implemented in SPM (fmripower.org; Mumford and Nichols, 2008). Test sample size was set at $n = 40$, which has been shown to be adequate for the unbiased detection of real effects in fMRI data (Yarkoni, 2009), and Type 1 error rate was set at $\alpha = 0.05$ for each ROI. The resulting power estimates therefore indicate the probability of detecting an effect at a statistical threshold of $P = 0.05$ in a future test sample of $n = 40$ subjects. Consistent with the established standards (Cohen, 1988), power values $> 80\%$ were considered as acceptable.

Effects of low-level kinematics. To identify and spatially localize the neural effects of low-level kinematics, we used frame-wise kinematic information (Roux *et al.*, 2013) of the triangles' physical movement properties, namely the 'instantaneous velocity' (defined separately for each triangle, V_{little} and V_{big}) and the 'relative distance' (*D*) between the triangles. The frame-wise kinematic regressors were i) down-sampled to 0.5 Hz (TR = 2 s), ii) z-transformed (mean 0 and s.d. 1), iii) convolved with the canonical HRF, iv) orthogonalized to the video clip regressors in order to remove the main effect of the animations and v) included as regressors of interest into the first-level (single-subject) models. MCQ ratings were modelled as well to facilitate comparisons against the implicit baseline. Parameter estimation followed the same GLM procedure as described above. Individual beta-images of each kinematic condition were subsequently subjected to a one-way analysis of variance for second-level statistical inference. Kinematic-sensitive brain regions were identified by testing voxel-wise beta values against the null hypothesis (H_0) of no increased response to any of the conditions using the minimum statistic compared to the global null (Nichols *et al.*, 2005). To lower the risk of false negatives (i.e. failure to detect an effect of kinematics in mentalizing-associated brain regions), significance was defined at a very liberal threshold of $P < 0.05$, uncorrected for multiple comparisons across the whole brain.

Results

Categorization performance

Mean overall categorization accuracy (calculated as percent correct) was 88.2 ± 7.5% (ToM: 97.8 ± 7.0%; GD: 75.0 ± 14.7%; R: 91.8 ± 14.8%). Emotional ratings of ToM video clips were consistently positive for two animations (coaxing: 0.91 ± 0.28 for the big and 0.81 ± 0.54 for the little triangle; surprising: 0.80 ± 0.59 for the big and 0.93 ± 0.33 for the little triangle), and neutral-to-negative for the remaining two animations (mocking: −0.49 ± 0.59 for the big and −0.10 ± 0.85 for the little triangle; seducing: −0.82 ± 0.48 for the big and 0.11 ± 0.93 for the little triangle; Figure 1).

Given the lower categorization accuracy for GD videos, supplemental analyses were conducted to explore the effects of classification performance on mentalizing-specific brain activation (see Supplementary Material).

Whole-brain responses to spontaneous mentalizing

We detected significant activation increases during the mentalizing condition ToM compared to the high-level baseline of GD behavior in areas that were repeatedly highlighted in prior ToM studies (for an overview, see Schurz *et al.*, 2014) and overlapped with the pre-defined social brain masks (see Table 1, Figure 3A). These included pSTS, anterior STS and TP, dmPFC, IFG and precuneus. In addition, increased activations were observed in a large cluster extending from the occipital lobe (including anterior and posterior portions of the lateral occipital complex, LO1 and LO2) to the fusiform and inferior temporal gyri.

Test–retest reliability and power

While paired *t*-tests did not reveal any significant changes in group-level brain activations from session 1 to session 2,

Table 1. Whole-brain activation during spontaneous mentalizing compared to agency perception (ToM > GD)

Region	*k*	x	y	z	t	P_{corr}
Medial temporal pole	1605	48	11	−35	12.00	<0.001
Middle temporal gyrus		51	−58	13	11.61	<0.001
Fusiform gyrus		42	−46	−17	9.58	<0.001
Superior temporal gyrus [Area PFm (IPL)]		63	−46	22	8.34	<0.001
Middle temporal gyrus	945	−54	−52	16	11.72	<0.001
Middle occipital gyrus [Area PGp (IPL)]		−39	−79	28	9.53	<0.001
Inferior occipital gyrus [hOc4lp (LO1)]	70	30	−94	1	11.47	<0.001
Cerebellum [Lobule VIIa crus 2]	189	−21	−79	−38	10.46	<0.001
Medial temporal pole	228	−42	14	−35	9.78	<0.001
Fusiform gyrus	217	−42	−49	−20	9.21	<0.001
Inferior occipital gyrus [hOc4lp (LO2)]		−42	−70	−8	7.00	<0.001
Cerebellum [Lobule VIIa crus 1]	145	24	−79	−35	8.96	<0.001
Precuneus	170	3	−55	43	8.8	<0.001
Middle occipital gyrus [hOc4la(LO1)]	63	−33	−94	−5	8.58	<0.001
IFG (pars orbitalis)	122	−45	29	−8	7.42	<0.001
IFG	69	57	29	7	6.99	<0.001
Middle frontal gyrus	51	−42	8	52	6.88	<0.001
Superior medial frontal gyrus	84	3	50	34	6.65	0.001
Cerebellum [Lobule IX]	13	−6	−55	−47	6.16	0.003
Supplementary motor area	15	9	14	67	6.01	0.004
Precentral gyrus	11	45	2	40	5.96	0.005
Parahippocampal gyrus [Amygdala (LB)]	4	24	−4	−23	5.71	0.011

Note: Cluster extent *k* is given at $P_{corr} < 0.05$, family wise error corrected for multiple comparisons across the whole brain. Regions were classified according to the Automated Anatomical Labeling Atlas (Tzourio-Mazoyer *et al.*, 2002). If applicable, anatomical labels were added in square brackets based on Anatomical Probability Maps (Anatomy toolbox; Eickhoff *et al.*, 2006). X-, y-, and z-coordinates MNI and statistical information refer to the peak voxel(s) in the corresponding cluster. *P* values are adjusted for family wise error correction for multiple comparisons across the whole brain.

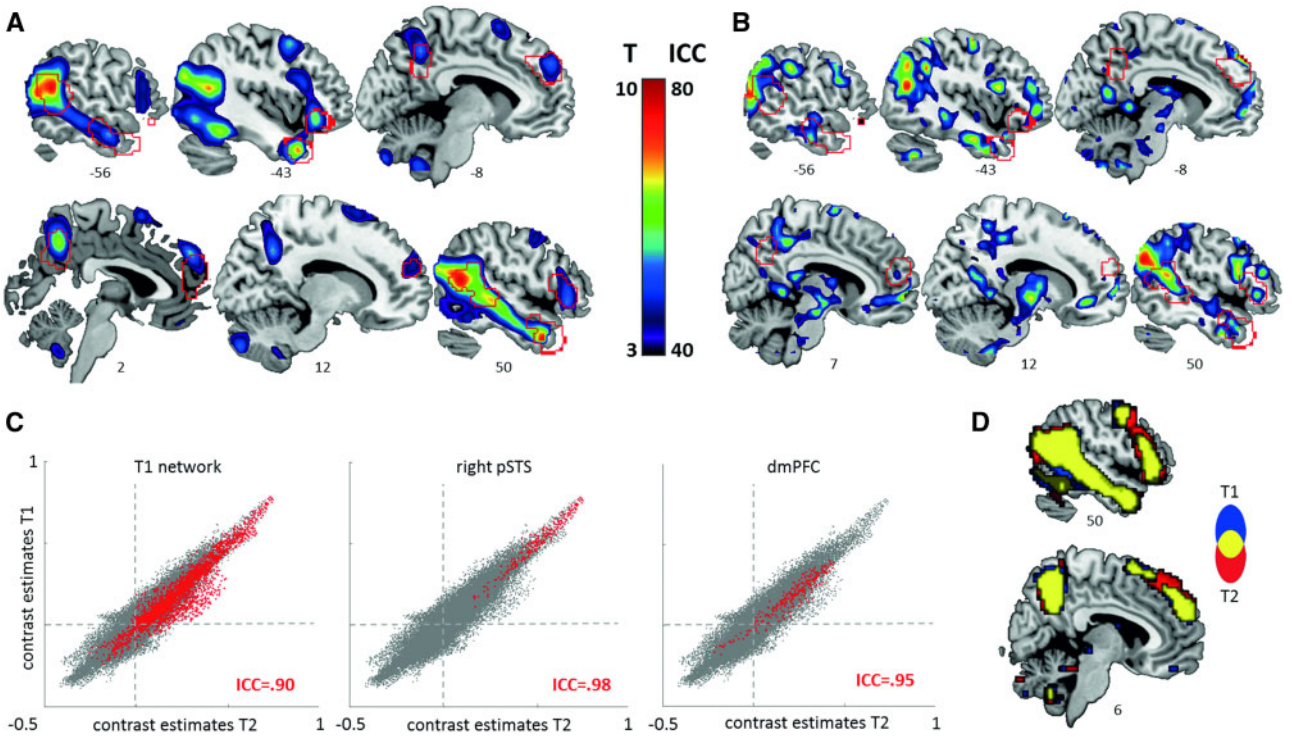

Fig. 3. Functional activation (A) and reliability metrics (B–D) during spontaneous mentalizing (ToM) compared to agency perception (GD). Sections display thresholded (A) T-maps and (B) ICC(3,1)-maps (i.e. single-voxel reliability). ROIs are outlined in red. (C) Spatial reliability of group activation maps is illustrated as scatter plots of voxel-wise contrast estimates at session 1 (T1) and session 2 (T2). Voxels belonging to the respective ROI are highlighted in red. Dashed lines designate zero on each axis. (D) Sections showing the overlap (in yellow) of whole-brain networks, defined at T1 (blue) and T2 (red) at a significance threshold of $P_{uncorr} < 0.001$.

reliabilities were heterogeneous across the different ROIs and metrics (see Table 2, Figure 3 B–D). Highest values were obtained for dice overlap and spatial reliability on the group level, indicating high cross-sectional reliability across ROIs. On the single-subject level, no ROI fell below the lower bound (0.40) of the defined range of acceptable ICC values for spatial reliability (calculated as the median of individual ICCs), suggesting that spatial activation patterns are sufficiently reliable in a longitudinal setting. Dice overlap for single-subject maps was more heterogeneous (maximum of 0.43 and 0.53 for the right pSTS at a threshold defined at $P < 0.001$ and $P < 0.005$, respectively). Despite this heterogeneity, the additional criterion of greater within- than between-subject overlap was met by most ROIs. Between-subject reliability as assessed as single-voxel reliability was also heterogeneous across ROIs (overall mean of 34% reliable voxels). A reliable portion of at least 10% of voxels was found for each ROI except for the precuneus. Larger portions were found in the pSTS, right TP, right IFG and left aMTG (47%–72%). Of note, the clustering of reliable voxels across the whole brain did not stringently overlap with the activation (Figure 3 A and B) or deactivation peaks (data not shown). Similar observations have been reported in previous reliability studies (Caceres *et al.*, 2009; Plichta *et al.*, 2012). Power calculations revealed that all ROIs performed well above the commonly adopted 80% threshold (Cohen, 1988).

Differential responses to low-level kinematics

Effects of low-level kinematic stimulus properties were mainly observed in areas of the dorsal visual pathway and the downstream oculomotor network (Goodale and Milner, 1992; Petit and Haxby, 1999), including the motion-sensitive extrastriate areas V3A and V5, superior parietal lobe (SPL) and frontal eye fields (FEF; at the caudal end of the superior frontal sulcus; Table 3 and Figure 4A). Visual comparison of whole-brain effects of low-level kinematics to both pre-defined ROIs (Figure 4B) and whole-brain observed (Figure 4C) mentalizing effects suggested largely non-overlapping activation patterns (no overlap for all social brain ROIs, below 6% for the T1 network at $P < 0.05$, uncorrected for multiple comparisons across the whole brain, see panels B and C in Figure 4).

Discussion

The human social brain is the focus of many neuroimaging studies dedicated to basic neuroscience and clinical research. Given the high clinical relevance of social dysfunction and the advance of large multicenter fMRI studies, empirical knowledge on the quality criteria of social tasks becomes increasingly important. To this end, the current study investigated the specificity, reliability and sensitivity of brain responses to the FHAs, a well-established set of experimental stimuli arising from cognitive psychology (Abell *et al.*, 2000). Among others, we demonstrate a strong differential engagement of key structures of the human social brain to stimuli engaging mentalizing processes relative to the high-level control stimuli challenging simple intention detection. We hope that our findings, discussed in more detail below, will be useful to guide the application of FHAs in ongoing large-scale studies aiming at identifying the neural mechanisms of normal and altered social cognition.

As our first main finding, we observed that the ToM video clips provoked a strong activation increase in brain regions that have been previously implicated in various aspects of social cognition. While the strongest responses mapped to

Table 2. Reliability and power of functional responses to spontaneous mentalizing compared to agency perception (ToM > GD)

	Session-wise activation: peak contrast value			Group map: cross-sectional reliability			Single subject map: longitudinal reliability					Single voxel: between-subject reliability	Power
	S1 (s.d.)	S2 (s.d.)	*P* value[a]: S1 vs S2	spatial ICC ±95% CI	overlap at $P<0.001$	overlap $P<0.005$	median spatial ICC (IQR)	mean within overlap at $P<0.001$ (s.d.)	*P* value[b]: within vs. between overlap at $P<0.001$	mean within overlap $P<0.005$ (s.d.)	*P* value[b]: within vs. between overlap at $P<0.005$	ICC > 0.40	in % (ES)
T1 network[c]	1.59 (0.56)	1.59 (0.47)	0.988	0.9 (0.90 0.91)	0.89	0.93	0.54 (0.41 0.65)	0.34 (0.22)	<0.001	0.42 (0.22)	<0.001	36%	99.97 (0.82)
pSTS R	1.15 (0.59)	1.15 (0.48)	0.995	0.98 (0.97 0.98)	1	1	0.69 (0.49 0.83)	0.43 (0.32)	0.008	0.53 (0.31)	0.016	72%	100 (1.22)
pSTS L	0.98 (0.39)	0.85 (0.32)	0.085	0.99 (0.98 0.99)	0.99	1	79 (0.65 0.82)	0.43 (0.33)	<0.001	0.52 (0.34)	0.002	43%	100 (1.18)
dmPFC	0.94 (0.61)	0.84 (0.41)	0.517	0.95 (0.94 0.96)	0.84	0.9	0.6 (0.40 0.74)	0.11 (0.17)	0.024	0.19 (0.21)	0.007	11%	83.79 (0.42)
preC	0.74 (0.35)	0.74 (0.29)	0.933	0.97 (0.96 0.97)	0.96	0.97	0.56 (0.24 0.79)	0.2 (0.31)	<0.001	0.25 (0.33)	0.002	6%	98.53 (0.62)
IFG R	0.75 (0.39)	0.87 (0.32)	0.163	0.96 (0.95 0.98)	0.85	0.92	0.66 (0.41 0.87)	0.25 (0.3)	0.016	0.36 (0.3)	0.01	48%	97.51 (0.58)
IFG L	0.77 (0.44)	0.91 (0.45)	0.282	0.89 (0.82 0.93)	0.92	0.93	0.61 (0.33 0.77)	0.16 (0.29)	0.043	0.25 (0.32)	0.097	14%	99.97 (0.81)
aMTG R	0.56 (0.26)	0.55 (0.27)	0.942	0.85 (0.78 0.90)	0.89	0.94	0.59 (0.45 0.73)	0.19 (0.28)	0.003	0.24 (0.33)	0.013	18%	99.95 (0.80)
aMTG L	0.41 (0.22)	0.36 (0.17)	0.351	0.73 (0.60 0.82)	0.7	0.8	0.45 (0.16 0.65)	0.06 (0.16)	0.048	0.14 (0.23)	0.014	47%	99.42 (0.67)
TP R	0.77 (0.25)	0.74 (0.31)	0.608	0.64 (0.53 0.72)	0.64	0.66	0.49 (0.18 0.63)	0.31 (0.3)	<0.001	0.36 (0.32)	<0.001	49%	100 (1.12)
TP L	0.8 (0.37)	0.65 (0.42)	0.147	0.4 (0.21 0.56)	0.26	0.43	0.4 (0.10 0.59)	0.05 (0.14)	0.161	0.11 (0.19)	0.024	33%	100 (1.06)

[a]P value as assessed by paired *t*-test on session-wise peak values,
[b]P value calculated as empirical *P* value from 10 000 repetitions (see Materials and methods),
[c]T1 network consisted of 8147 voxels
S1: session 1, S2: session 2, ICC: intraclass correlation coefficient, R_{over}: spatial overlap measure, CI: confidence interval, s.d.: standard deviation, IQR: interquartile range (Q1 Q3), ES: effect size (calculated as Cohen's d), pSTS: posterior temporal sulcus, dmPFC: medial prefrontal cortex, preC: precuneus, IFG: inferior frontal gyrus: aMTG, anterior middle temporal gyrus, TP: temporal pole, L: left, R: right.

Table 3. Whole-brain activation to low-level kinematics

Region	k	x	Y	Z	t	P_{corr}
Middle occipital gyrus [hOc4lp (LO2), hOc5 (V5/MT)]	575	−45	−79	13	6.19	<0.001
Superior occipital gyrus [hOc4d (V3A)]		−21	−88	40	2.34	
Superior parietal lobule [Area 7A (SPL)]	1376	15	−61	67	5.32	<0.001
Superior parietal lobule [Area 5L (SPL)]		−18	−55	67	4.11	
Middle temporal gyrus [hOc4lp (LO2), hOc5 (V5/MT)]	262	45	−73	7	5.08	<0.001
Lingual gyrus [hOc1 (V1), hOc2 (V2)]	68	9	−91	−8	3.70	<0.001
Precentral gyrus (FEF, BA 6)	327	24	−13	58	3.61	<0.001
Superior frontal gyrus (FEF, BA 6)	370	−21	−10	55	3.22	<0.001
Cerebellum [Lobule VIIIa]	19	−30	−40	−47	1.45	0.983
Cerebellum [Lobule VIIIa]	16	33	−40	−44	1.02	1.000
Rectal gyrus	4	9	47	−20	0.97	1.000
Medial temporal pole	1	39	20	−38	0.35	1.000

Note: Cluster extent k is given at $P < 0.05$ (uncorrected for multiple comparisons across the whole brain), BA, Brodmann Area, FEF, frontal eye field. Regions were classified according to the Automated Anatomical Labeling Atlas (Tzourio-Mazoyer *et al.*, 2002). If applicable, anatomical labels were added in square brackets based on Anatomical Probability Maps (Anatomy toolbox; Eickhoff *et al.*, 2006). *X*-, *y*-, and *z*-coordinates MNI and statistical information refer to the peak voxel(s) in the corresponding cluster. *P* values are adjusted for family wise error correction for multiple comparisons across the whole brain. Note that all clusters were identfied at a liberal voxel-wise threshold of of $P < 0.05$, uncorrected for multiple comparisons across the whole brain, in order to maximize sensitivity for potential confounding effects.

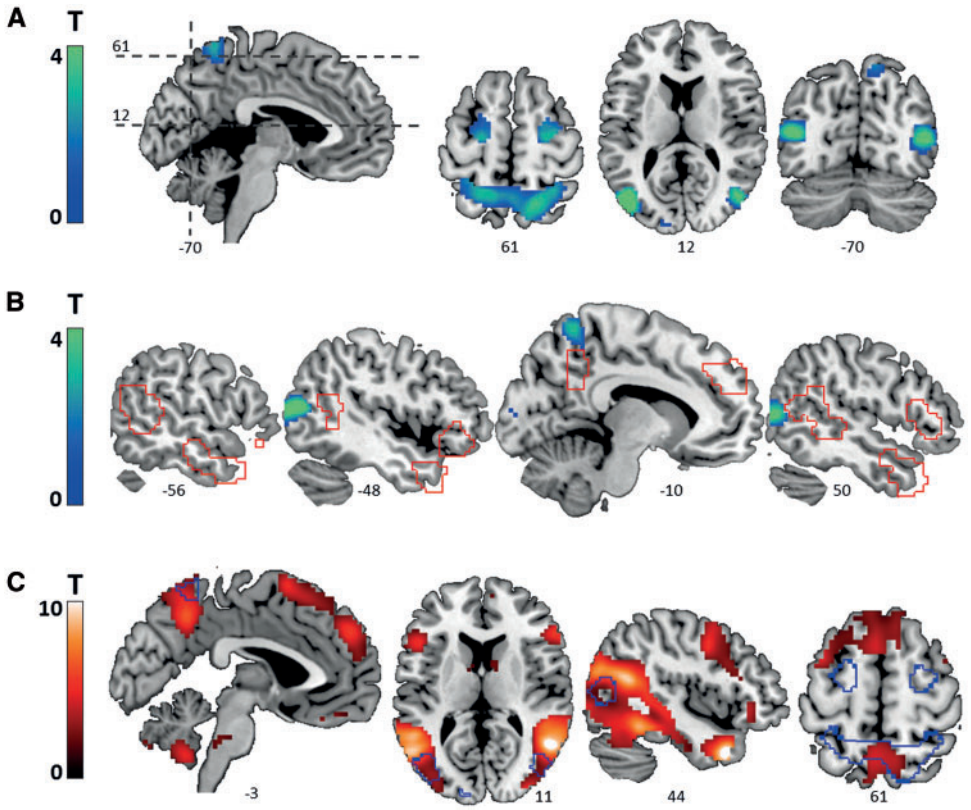

Fig. 4. Whole-brain effects of low-level kinematic stimulus properties. (A) Sections displaying activated clusters. (B) Sections showing outlines of the pre-defined mentalizing network (red), overlaid on whole-brain effects of low-level kinematics. (C) Outlines of the whole-brain effects of low-level kinematics are projected on sections displaying whole-brain effects of mentalizing (ToM > GD). All statistical maps were thresholded at $P < 0.05$, uncorrected for multiple comparisons across the whole brain to minimize false negatives.

occipito-temporal areas involved in social perception (Allison *et al.*, 2000; Deen *et al.*, 2015), the activations of the inferior parietal lobule and IFG are consistent with the self-referential representation of the observed actions (e.g. mirroring; Iacoboni and Dapretto, 2006). In addition, the ToM animations recruited structures known for subserving cognitive functions that allow for the higher-order representation of complex social scenes, such as perspective taking (precuneus; Cavanna and Trimble, 2006), abstract reasoning (dmPFC; Bzdok *et al.*, 2013), and social knowledge retrieval (TP; Olson *et al.*, 2013). The observed breadth of activations is hereby in line with current network accounts on social cognition, which propose that specialized but highly interrelated neural circuits operate on an implicit-to-explicit continuum (Yang *et al.*, 2015). Applied to our findings, this framework would suggest that the strong engagement of social-perceptive areas reflects the implicit bottom-up demands

of spontaneous mentalizing. Conversely, the engagement of higher-order areas likely reflects the more explicit demands of the stimuli, for example, those related to the evaluation and rating of the ToM video clips (MCQ-cat, MCQ-feelings). We thus conclude from these data that, albeit their minimal task demands, the FHAs are suitable for challenging both bottom-up sensory and higher-order representational areas and processes in the social brain. Moreover, in contrast to purely passive tasks, the implemented non-verbal ratings provide a form of behavioral control during data acquisition (e.g. to secure that the subjects indeed performed the task and understood the presented social scenarios). For instance, the observed category ratings imply that the subjects recognized the different levels of social significance of the presented interaction. Of note, our supplemental analyses suggest that misclassifications can meaningfully explain activation within the mentalizing network (see Supplementary Material). In addition, the valence ratings suggest that the emotional significance (or leitmotif) of the cover stories was generally understood, with a certain variability across subjects that may point to (potentially clinically and sub-clinically meaningful) differences in the acquisition of the social emotional concepts of the animations. We expect these behavioral measures to be very useful in future studies since they may, for example, guide the identification of the neural correlates of individual differences in subjective stimulus evaluation and allow for the comparison of performance-adjusted groups.

As a second study goal, we examined the functional correlates of the low-level kinematics of the FHAs. Prior work identified kinematic differences in form of higher-immobilization rate and lower relative distance of the animated shapes in the ToM conditions, which related to differences in the eye-movement towards the FHAs conditions (Roux *et al.*, 2013). With respect to neural network activation, our data suggests that the predominant effect of the stimulus kinematics is rather localized and is confined to the visuo-oculumotor circuitry (Petit and Haxby, 1999). Notably, differences in the kinematic profile are an inherent determinant of the social significance of the FHAs (and plausibly also of natural social interactions) and are therefore not fully separable from the video clip conditions (Scholl and Tremoulet, 2000; Roux *et al.*, 2013). However, our results suggest that their relevance to higher order social brain responses to the task are rather low. Importantly, the kinematic-specific responses did not even overlap with the social condition-specific effects within visual areas. The latter were observed in the shape-selective area LO1 as well as in the mid-fusiform gyrus, both representing higher order visual areas involved in the detection of agency and social meaning (Shultz and McCarthy, 2014; Malikovic *et al.*, 2015).

Besides the specificity of the neural responses, we studied the task's capability to robustly activate the targeted system by means of test–retest reliability assessments along with power calculations of the elicited neural signals. In principle, a reliable task allows the attribution of detected signals to factors other than unstable task effects. The reproducibility of fMRI data can be compromised by multiple factors, such as changes in acquisition parameters (e.g. scanner hardware, field strength, image signal-to-noise ratio) and subject-related factors (e.g. habituation, motion, cognitive strategies; Raemaekers *et al.*, 2007; Caceres *et al.*, 2009; Bennett and Miller, 2010; Gorgolewski *et al.*, 2013). Prior studies have assessed the reliability of brain responses in various domains, including emotion and motivation (e.g. Johnstone *et al.*, 2005; Plichta *et al.*, 2012, 2014; Lipp *et al.*, 2014), executive (e.g. Caceres *et al.*, 2009; Plichta *et al.*, 2012) and sensorimotor functions (e.g. Zandbelt *et al.*, 2008; Caceres *et al.*, 2009), but not for mentalizing tasks targeting the social brain. Notably, unlike many prior reliability studies, we specifically focused on the robustness of the high-level experimental contrast, controlled for a range of basic physiological confounds, and aimed at maximizing the generalizability of results by constraining our analysis to a set of meta-analytically derived ROIs.

We applied two different metrics, dice overlap and spatial reliability, to the group- and single-subject data in order to quantify aspects of cross-sectional and longitudinal reliability, respectively. These analyses revealed good-to-excellent reliability of group-level responses to FHAs across ROIs and metrics, which is in line with the observation of a generally high reproducibility of other robust fMRI paradigms studied cross-sectionally (Plichta *et al.*, 2012). Longitudinal reliability was lower, in particular for the spatial overlap measure, although the criterion of greater within- than between-overlap of suprathreshold activation (Gorgolewski *et al.*, 2013) within predefined ROIs was met by most areas of the social brain. Spatial reliability was fair-to-good for all ROIs, and even excellent for the right pSTS. These results are in line with previous reports of lower reliability of fMRI results obtained on the single-subject compared to the group level (Raemaekers *et al.*, 2007; Plichta *et al.*, 2012; Lipp *et al.*, 2014).

Across ROIs, the overall reliability pattern suggests that activity was less reproducible in areas linked to higher-order cognitive processes, such as meta-cognition and self-reflection (e.g. dmPFC, precuneus; Cavanna and Trimble, 2006; Bzdok *et al.*, 2013). A plausible reason for this observation is the fact that the triangle animations allow for a certain degree of 'cognitive freedom' (Gorgolewski *et al.*, 2013) since they allow to be interpreted differently by different subjects and might therefore be associated with a greater variability of neural responses. The reliability of the pSTS, in contrast, was excellent across levels (i.e. single-subject and group level) and metrics (i.e. spatial reliability and dice overlap), which is in line with this region's involvement in more implicit, bottom-up and thus possibly more invariant processes linked to social perception (Deen *et al.*, 2015). The repeated exposure of FHAs stimuli within a two-week interval might therefore not influence bottom-up processes responsible for triggering spontaneous mentalizing, but may have an impact on how the videos are interpreted upon second presentation (e.g. no need for a de-novo reconstruction of the presented social scenario). In more general terms, our finding of different reliabilities across ROIs likely reflects the difference between more constrained bottom-up and more flexible top-down processing of social information.

The right pSTS also yielded the best differentiation between subjects, with more than 70% of voxels displaying at least fair-to-good between-subject reliability (i.e. single-voxel reliability with $ICC(3,1) > 0.40$). Since greater heterogeneity between subjects leads to higher ICC values, our results suggest that pSTS activity, besides being highly consistent across sessions, is sufficiently heterogeneous across subjects and is thus particularly suitable for endophenotype research (Raemaekers *et al.*, 2007).

Besides test–retest reliability, statistical sensitivity was assessed using power calculations for a future test sample of 40 subjects. Here, all key regions of the social brain showed a very good probability to detect a functional effect while keeping the type I error rate below 5%. The good statistical sensitivity may be a consequence of the bottom-up driven processes of FHAs, but also the robust block structure of the task and the higher signal-to-noise ratios which result from the chosen voxel size (Bennett and Miller, 2010) and smoothing kernel (Caceres *et al.*, 2009).

Following the systems biology approach, the aim of task-based fMRI in biomarker research is to access specific functional networks with sufficient reliability and sensitivity in order to identify markers of normal or pathogenic processes or of treatment response (Biomarkers Definitions Working Group, 2001). While the obtained fMRI phenotypes may, to some extent, be specific for the employed task (e.g. spontaneous mentalizing vs. false-belief reasoning) and associated behavior, they reflect the interplay of large-scale and oftentimes overlapping neural circuitries. Aberrant activation patterns in patient groups can therefore point to a disruption of information processing within or between these networks. According to this rationale, we believe that our neuroimaging findings support the following conclusions and recommendations for biomarker research on human mentalizing processes:

1. The FHAs lead to a relatively specific activation of social brain areas including networks involved in social perception, action observation and ToM. The high-level contrast of the task allows for the separation of effects of mentalizing from those related to simple intention attribution. Although the observed condition-dependent differences in activation appear to be mainly quantitative, qualitative differences might emerge in clinical populations such as patients with ASD (Castelli *et al.*, 2002).
2. Within our fMRI paradigm and the associated data acquisition and analysis procedures, the FHAs allow for a reliable characterization of social brain activations in cross-sectional and—to a certain extent—also longitudinal study designs. These stimuli therefore represent a promising means to identify and stratify imaging biomarkers and validate pharmacological interventions (Loth *et al.*, 2015), although their use in clinical populations additionally requires to demonstrate reliably altered brain responses in these samples.
3. The pSTS additionally demonstrated high between-subject reliability and is therefore particularly suited for tracing neural correlates of inter-individual differences, such as learning-related differences or potential genetically influenced neural traits (Raemaekers *et al.*, 2007). For instance, the repeated observation of pSTS dysfunctions in highly heritable disorders such as ASD (e.g. Zilbovicius *et al.*, 2006) supports the proposed value of this region for genetic research.

Finally, the specific and reliable brain responses to FHAs encourage the evaluation and use of this task for the characterization of mentalizing-specific functional connectivity differences in clinical research. We expect this line of research to offer insights into preserved and altered human social brain network dynamics in health and disorders with prominent social dysfunction such as schizophrenia and autism.

Acknowledgements

We thank Urs Braun, Emanuel Schwarz and Ceren Akdeniz for valuable input on the manuscript and Dagmar Gass for research assistance.

Funding

This work was supported by the European Community's Seventh Framework Programme under the grant agreements No. 115300 (Project EU-AIMS), No. 602805 (Project EU-AGGRESSOTYPE), No. 602450 (Project EU-IMAGEMEND), the German Federal Ministry of Education and Research [grant No. 01ZX1314GM (Project IntegraMent); grant No. 01GQ1102 to H.T.], the Agence Nationale de la Recherche (ANR-09-BLAN-0327, ANR-11-IDEX-0001-02 PSL* and ANR-10-LABX-0087 to P.R.) and Assistance Publique – Hôpitaux de Paris–Centre National de la Recherche Scientifique (APHP–CNRS to P.R.).

Supplementary data

Supplementary data are available at *SCAN* online.

Conflict of interest. None declared.

References

Abell, F., Happé, F., Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cognitive Development*, **15**, 1–16.

Adolphs, R. (2009). The social brain: neural basis of social knowledge. *Annual Review of Psychology*, **60**, 693–716.

Allison, T., Puce, A., McCarthy, G. (2000). Social perception from visual cues: role of the STS region. *Trends in Cognitive Sciences*, **4**(7), 267–78.

Apperly, I.A., Butterfill, S.A. (2009). Do humans have two systems to track beliefs and belief-like states? *Psychological Review*, **116**(4), 953–70.

Bennett, C.M., Miller, M.B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences*, **1191**, 133–55.

Biomarkers Definitions Working Group. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics*, **69**(3), 89–95.

Bzdok, D., Langner, R., Schilbach, L., *et al.* (2013). Segregation of the human medial prefrontal cortex in social cognition. *Frontiers in Human Neuroscience*, **7**, 232.

Caceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C., Mehta, M.A. (2009). Measuring fMRI reliability with the intra-class correlation coefficient. *Neuroimage* **45**(3), 758–68.

Castelli, F., Frith, C., Happé, F., Frith, U. (2002). Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain* **125**(Pt 8), 1839–49.

Castelli, F., Happé, F., Frith, U., Frith, C. (2000). Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *Neuroimage* **12**(3), 314–25.

Cavanna, A.E., Trimble, M.R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain* **129**(Pt 3), 564–83.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Revised Edition. Hillsdale, NJ: Erlbaum.

Deen, B., Koldewyn, K., Kanwisher, N., Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. *Cerebral Cortex*, **25**(11), 4596–609.

Eickhoff, S.B., Heim, S., Zilles, K., Amunts, K. (2006). Testing anatomically specified hypotheses in functional imaging using cytoarchitectonic maps. *Neuroimage*, **32**(2), 570–82.

Fleiss, J.L. (1986). *The Design and Analysis of Clinical Experiements*. New York, NY: Wiley.

Frith, U. (2004). Emanuel Miller lecture: confusions and controversies about Asperger syndrome. *Journal of Child Psychology and Psychiatry*, **45**(4), 672–86.

Goodale, M.A., Milner, A.D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, **15**(1), 20–5.

Gorgolewski, K.J., Storkey, A.J., Bastin, M.E., Whittle, I., Pernet, C. (2013). Single subject fMRI test-retest reliability metrics and confounding factors. *Neuroimage*, **69**, 231–43.

Iacoboni, M., Dapretto, M. (2006). The mirror neuron system and the consequences of its dysfunction. *Nature Reviews Neuroscience*, **7**(12), 942–51.

Johnstone, T., Somerville, L.H., Alexander, A.L., *et al.* (2005). Stability of amygdala BOLD response to fearful faces over multiple scan sessions. *Neuroimage*, **25**(4), 1112–23.

Lipp, I., Murphy, K., Wise, R.G., Caseras, X. (2014). Understanding the contribution of neural and physiological signal variation to the low repeatability of emotion-induced BOLD responses. *Neuroimage*, **86**, 335–42.

Loth, E., Spooren, W., Ham, L.M., *et al.* (2015). Identification and validation of biomarkers for autism spectrum disorders. *Nature Reviews Drug Discovery*, **15**(1), 70–3.

Malikovic, A., Amunts, K., Schleicher, A., *et al.* (2015). Cytoarchitecture of the human lateral occipital cortex: mapping of two extrastriate areas hOc4la and hOc4lp. *Brain Structure and Function*, **221**(4), 1877–97.

Mar, R.A. (2011). The neural bases of social cognition and story comprehension. *Annual Review of Psychology*, **62**, 103–34.

Mar, R.A., Macrae, C.N. (2007). Triggering the intentional stance. *Novartis Foundation Symposium*, **278**, 111–20. discussion 20-33, 216-21.

Mumford, J.A., Nichols, T.E. (2008). Power calculation for group fMRI studies accounting for arbitrary design and temporal autocorrelation. *Neuroimage*, **39**(1), 261–8.

Murphy, D., Spooren, W. (2012). EU-AIMS: a boost to autism research. *Nature Reviews Drug Discovery*, **11**(11), 815–6.

Nichols, T., Brett, M., Andersson, J., Wager, T., Poline, J.B. (2005). Valid conjunction inference with the minimum statistic. *Neuroimage*, **25**(3), 653–60.

Olson, I.R., McCoy, D., Klobusicky, E., Ross, L.A. (2013). Social cognition and the anterior temporal lobes: a review and theoretical framework. *Social Cognitive and Affective Neuroscience*, **8**(2), 123–33.

Petit, L., Haxby, J.V. (1999). Functional anatomy of pursuit eye movements in humans as revealed by fMRI. *Journal of Neurophysiology*, **82**(1), 463–71.

Plichta, M.M., Grimm, O., Morgen, K., *et al.* (2014). Amygdala habituation: a reliable fMRI phenotype. *Neuroimage*, **103C**, 383–90.

Plichta, M.M., Schwarz, A.J., Grimm, O., *et al.* (2012). Test-retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery. *Neuroimage*, **60**(3), 1746–58.

Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R.J., Kahn, R.S., Ramsey, N.F. (2007). Test-retest reliability of fMRI activation during prosaccades and antisaccades. *Neuroimage*, **36**(3), 532–42.

Rombouts, S.A., Barkhof, F., Hoogenraad, F.G., Sprenger, M., Scheltens, P. (1998). Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. *Magnetic Resonance Imaging*, **16**(2), 105–13.

Roux, P., Passerieux, C., Ramus, F. (2013). Kinematics matters: a new eye-tracking investigation of animated triangles. *Quarterly Journal of Experimental Psychology (Hove)*, **66**(2), 229–44.

Schneider, D., Slaughter, V.P., Bayliss, A.P., Dux, P.E. (2013). A temporally sustained implicit theory of mind deficit in autism spectrum disorders. *Cognition*, **129**(2), 410–7.

Scholl, B.J., Tremoulet, P.D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, **4**(8), 299–309.

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., Perner, J. (2014). Fractionating theory of mind: a meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, **42**, 9–34.

Senju, A., Southgate, V., White, S., Frith, U. (2009). Mindblind eyes: an absence of spontaneous theory of mind in Asperger syndrome. *Science*, **325**(5942), 883–5.

Shrout, P.E., Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, **86**(2), 420–8.

Shultz, S., McCarthy, G. (2014). Perceived animacy influences the processing of human-like surface features in the fusiform gyrus. *Neuropsychologia*, **60**, 115–20.

Smith, S.M., Beckmann, C.F., Ramnani, N., *et al.* (2005). Variability in fMRI: a re-examination of inter-session differences. *Human Brain Mapping*, **24**(3), 248–57.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., *et al.* (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, **15**(1), 273–89.

White, S.J., Coniston, D., Rogers, R., Frith, U. (2011). Developing the Frith-Happé animations: a quick and objective test of Theory of Mind for adults with autism. *Autism Research*, **4**(2), 149–54.

Yang, D.Y., Rosenblau, G., Keifer, C., Pelphrey, K.A. (2015). An integrative neural model of social perception, action observation, and theory of mind. *Neuroscience and Biobehavioral Reviews*, **51**, 263–75.

Yarkoni, T. (2009). Big correlations in little studies: inflated fMRI correlations reflect low statistical power-commentary on Vul et al. (2009). *Perspectives on Psychological Science*, **4**(3), 294–8.

Zandbelt, B.B., Gladwin, T.E., Raemaekers, M., *et al.* (2008). Within-subject variation in BOLD-fMRI signal changes across repeated measurements: quantification and implications for sample size. *Neuroimage*, **42**(1), 196–206.

Zilbovicius, M., Meresse, I., Chabane, N., Brunelle, F., Samson, Y., Boddaert, N. (2006). Autism, the superior temporal sulcus and social perception. *Trends in Neurosciences*, **29**(7), 359–66.