NeuroImage 161 (2017) 9–18

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs

Jorie Koster-Hale [a,1], Hilary Richardson [a,*,1], Natalia Velez [b], Mika Asaba [b], Liane Young [c], Rebecca Saxe [a]

[a] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[b] Department of Psychology, Stanford University, Stanford, CA 94305, USA
[c] Department of Psychology, Boston College, Chestnut Hill, MA 02467, USA

ARTICLE INFO

Keywords:
Theory of mind
fMRI
Multi-voxel pattern analysis (MVPA)

ABSTRACT

The human capacity to reason about others' minds includes making causal inferences about intentions, beliefs, values, and goals. Previous fMRI research has suggested that a network of brain regions, including bilateral temporo-parietal junction (TPJ), superior temporal sulcus (STS), and medial prefrontal-cortex (MPFC), are reliably recruited for mental state reasoning. Here, in two fMRI experiments, we investigate the representational content of these regions. Building on existing computational and neural evidence, we hypothesized that social brain regions contain at least two functionally and spatially distinct components: one that represents information related to others' motivations and values, and another that represents information about others' beliefs and knowledge. Using multi-voxel pattern analysis, we find evidence that motivational versus epistemic features are independently represented by theory of mind (ToM) regions: RTPJ contains information about the justification of the belief, bilateral TPJ represents the modality of the source of knowledge, and VMPFC represents the valence of the resulting emotion. These representations are found only in regions implicated in social cognition and predict behavioral responses at the level of single items. We argue that cortical regions implicated in mental state inference contain complementary, but distinct, representations of epistemic and motivational features of others' beliefs, and that, mirroring the processes observed in sensory systems, social stimuli are represented in distinct and distributed formats across the human brain.

1. Introduction

Successful social interaction requires reasoning about the minds of other people: observing not just how someone is behaving, but inferring *why* they are behaving that way. By considering mental states – desires, values, beliefs, and expectations – we can predict, explain, and evaluate each other's actions, building a "theory of mind" (ToM). How does the human brain support these complex, fast, and often spontaneous social inferences? Existing neuroimaging research suggests where to look: a group of brain regions is robustly and reliably recruited when participants consider the minds of other people, including temporal parietal junction (RTPJ, LTPJ), right superior temporal sulcus (RSTS), precuneus (PC), and medial prefrontal cortex (MPFC) (Saxe and Powell, 2006) (for reviews see (Carrington and Bailey, 2009; Frith and Frith, 2012)). While earlier research has focused on localizing regions selectively involved in mental state reasoning, a critical open challenge is to systematically characterize the processes that are supported by different components of this system.

One powerful way to probe neural processes is to ask how the representation of a stimulus is reformatted at each stage of neural computation: different populations of neurons within a network may contribute to a common task by representing different features or aspects of the same stimulus or task. In the ventral visual stream, stimuli are represented in distinct and distributed formats across regions; e.g., while one neural population represents an item's color, a distinct population represents its shape (DiCarlo et al., 2012; Kamitani and Tong, 2005; Kourtzi, 2001; Lafer-Sousa and Conway, 2013; Tanaka, 1993). Asking which features of a stimulus can be linearly decoded from each population of neurons can reveal the kinds of representations that those populations support – if a region distinguishes between blue and red stimuli, it likely

* Corresponding author. 43 Vassar Street, 46-4021, Cambridge, MA 02139, USA.
E-mail address: hlrich@mit.edu (H. Richardson).
[1] Shared first authorship.

http://dx.doi.org/10.1016/j.neuroimage.2017.08.026
Received 7 March 2017; Accepted 8 August 2017
Available online 12 August 2017

represents color; if it distinguishes between squares and circles, it likely is sensitive to shape. Thus, this approach can inform our understanding of both what specific neural populations within a network are doing, and the processes engaged by the network as a whole.

Here, we leveraged this approach to probe the representational architecture of mental state inference. Building on existing computational and neural evidence, we hypothesized that theory of mind in the human brain contains at least two functionally and spatially distinct components: one that represents information related to others' motivations and values, and another that represents others' epistemic states – evaluating the source of their beliefs and knowledge. State of the art computational models of intuitive ToM suggest that these two components are sufficient to accurately model basic social behavior, using (i) a model of planning, to predict the action an agent will choose given their goals, desires, values, costs, and beliefs; and (ii) a model of belief formation, to predict the beliefs the agent will form given their perceptual access and inferential processes (Baker et al., 2017, 2009; Bello, 2012; Jara-Ettinger et al., 2012).

Moreover, multiple lines of evidence suggest regions in the ToM network may be differentially sensitive to tasks and stimuli related to motivational versus epistemic reasoning. Previous work examining overall activity of different regions finds that lateral regions (right and left TPJ) are recruited more by stimuli and tasks requiring reasoning about others' beliefs and intentions (epistemic or "cognitive" ToM (Schlaffke et al., 2014; Schnell et al., 2011)), while medial PFC is recruited more by inferences about emotions and preferences (motivational or "affective" ToM (Amodio and Frith, 2006; Etkin et al., 2011; Hynes et al., 2006; Sebastian et al., 2012; Shamay-Tsoory and Aharon-Peretz, 2007; Shamay-Tsoory et al., 2006; Leopold et al., 2011; Shamay-Tsoory, 2011)).

We hypothesized that the differential activation observed in these regions reflect a division in the types of computational processes supported by these regions – a socially relevant stimulus will be represented in RTPJ in an epistemic feature space, while the same stimulus will be represented in MPFC in a motivational feature space, each region collapsing across some dimensions of the stimuli, while emphasizing others. This hypothesis is supported by growing evidence suggesting that the population-level activity in MPFC contains abstract, multimodal information relevant to the motivational component of ToM. Systematic patterns of activity in MPFC are evoked when observing another individual (a) make a positive versus negative dynamic facial expression (Harry et al., 2013; Peelen et al., 2010; Said, 2010; Said et al., 2010), (b) make a positive versus negative vocal expression (Peelen et al., 2010), (c) succeed versus fail to complete a goal (like throwing a ball into a net) (Skerry and Saxe, 2014), or (d) get included in versus excluded from a social group (Skerry and Saxe, 2014). How pleasant the experience is for the protagonist (i.e., the valence of the experience) best explains the pattern of response in MPFC to verbal descriptions of 200 unique emotional events (Skerry and Saxe, 2015).

However, a key open question is whether there is neural evidence for an epistemic component of ToM (reasoning about the epistemology of others' beliefs) that is distinct from motivational representations. Earlier work – leveraging previously collected data designed to ask orthogonal questions – hints that relevant epistemic information about others' beliefs may be represented by population-level activity in TPJ. For example, bilateral TPJ contains information about the perceptual source of another person's knowledge: whether beliefs were formed based on visual or auditory evidence (Koster-Hale et al., 2014). Moreover, TPJ tracks whether a harmful action was taken with full foreknowledge versus in ignorance (Koster-Hale et al., 2013). Here, we directly test the hypothesis that TPJ supports processes related to this kind of social epistemic reasoning.

Building on this previous work, this study examines whether social stimuli, like sensory stimuli, are reformatted into distinct and complementary representations across the human brain, by testing whether processes supporting motivational and epistemic components of ToM are functionally distinct, and whether these processes are reflected in neural populations in MPFC versus TPJ, respectively. We presented verbal narratives describing others' mental states, using minor changes in wording to simultaneously manipulate features relevant for evaluating beliefs and predicting emotions. First, replicating previous work, we manipulated whether the evidence for a belief was visual or auditory (the modality of the evidence). Second, in a previously untested manipulation, we changed whether the agent's evidence was strongly or weakly supportive of the belief (the justification of the belief). Evidence justification provides a particularly strong test for features of intuitive epistemology because it is abstract (rather than tied to specific sensory features), context specific (what might be good evidence for one conclusion could be poor evidence for another), and directly related to reasoning about the minds of others (determining whether the agent is a reliable, rational informant (Kovera et al., 1991; Miene et al., 1993; Olson, 2003)). Third, we manipulated whether the agent heard evidence first-hand versus via hearsay (reported from another character), as a test of whether the directness of evidence is represented in social brain regions. This manipulation was motivated by work on testimony and communicative inference (Kovera et al., 1991; Miene et al., 1993; Olson, 2003), developmental research on trust and testimony (Clément et al., 2004; Koenig and Harris, 2005; Lucas et al., 2013; Robinson et al., 2008), and evidence that acquisition of evidential markings that indicate source of knowledge is related to source monitoring and theory of mind development (Papafragou and Li, 2001) (Papafragou et al., 2007) (Ozturk and Papafragou, 2016; Ünal and Papafragou, 2016). Finally, we manipulated the emotional valence (positive or negative) of the main character's final state. Using patterns of activity in independently identified regions of interest (ROIs), we tested whether we could classify the same items according to these distinct features differentially across regions. Specifically, we hypothesized that epistemic features of others' beliefs (justification, source modality) would be represented in bilateral TPJ, and that motivational features (valence) would be represented in (D/M/V) MPFC. In order to test for the specificity of effects to these a priori hypothesized regions, we conducted the same tests in other ToM regions as well as in a set of control regions involved in language processing.

2. Materials and methods

2.1. Experiment summary

Participants in the scanner listened to short narratives in which a protagonist came to hold a belief based on evidence that varied in justification, modality, directness, and valence (Fig. 1). After each story, participants judged whether the protagonist felt happy or sad; half of all stories ended happily. Stories appeared in all four epistemic conditions across participants, crossing modality and justification, and were matched in both low-level (e.g. length, reading ease, lexical frequency) and high-level features (e.g. background and conclusion content); each participant heard every story in one of the four conditions. This design maximizes variance within condition (each stimulus was a different story) while minimizing differences across conditions (matched stories varied only in the words describing inputs to the character's belief formation, described in the *evidence* section of the stimulus). Thus, successful classification, which requires training and testing on distinct stories (rather than repeated instances of the same story), provides evidence for an abstract representation of the manipulated dimension that generalizes across heterogeneous inputs. In a second experiment, we conduct a conceptual replication in order to test the robustness of representations of the justification of others' beliefs, in the context of moral judgments.

2.2. Participants

All participants were native English speakers, had normal or

	Seeing Strong Evidence (SS)	Seeing Weak Evidence (SW)	Hearing First-person (HF)	Hearing Hearsay (HS)
Background (16s)	Julie had wanted a puppy of her very own for years. Every holiday, she asked her parents to give her a puppy, but they had always said that they couldn't afford it. On Julie's ninth birthday, she woke up and ran downstairs to her parents.			
Evidence (4-16s)	In the **bright kitchen**, she **saw** a **newly installed doggy door**.	In the **cluttered kitchen**, she **saw** a **few small toys on the floor.**	In the **quiet kitchen**, she **heard** a **quiet, excited yipping**.	In the kitchen, **her mom told her** that there was a **newly installed doggy door**.
Conclusion (2-4s)	Julie asked her mom, "Did we get a new puppy?!"			
Task	Happy or Sad?			

Fig. 1. Example stimuli from Exp. 1. In the fMRI task, participants listened to audio recordings of 40 target stories and 10 physical control stories. Stories lasted from 22 to 36 s. Individual participants saw each target story in one of the four epistemic conditions, forming a matched and counterbalanced design; every target story occurred in all four conditions across participants. After each story, participants heard the question "Happy or sad?" and indicated whether the main character in the story felt happy or sad using a button press (left/right).

corrected-to-normal hearing and vision, and gave written informed consent in accordance with the requirements of Institutional Review Board at MIT. All participants were recruited from campus and the surrounding Boston area and received payment for participation.

In **Exp. 1**, 20 right-handed adults participated (11 women; mean age ± SD, 28.5 years ± 6.8). In **Exp. 2**, 20 right-handed adults (12 women; 21.25 ± 2.6) participated. The participant samples for Exp. 1 and Exp. 2 were independent. Two participants in Exp. 1 were dropped for excessive head motion, and one failed to complete the experiment, leaving 17 in all analyses; no participants were excluded from Exp. 2.

2.3. Stimuli

2.3.1. Experiment 1

Participants were scanned while listening to audio recordings of 40 stories about a character's beliefs and emotions and 10 physical control stories (Fig. 1 and Supplementary Materials (SM)). Story texts are available for download (http://saxelab.mit.edu/stimuli.php). The stories were digitally recorded by 11 female speakers at a sampling rate of 44,100 to produce 32-bit digital sound files. Each story was presented in 4 sections: (i) background information (identical across conditions, 16 s, 32 words ± 2.5), (ii) evidence (6–14 s; 25 words ± 2.3), (iii) conclusion (2–4 s, 10 words ± 2.4), and (iv) participant response (6 s). The four epistemic conditions were distinguished only in the "evidence" section.

In **Seeing Strong Evidence** (SS) stories, the protagonist makes an inference based on clear visual access to reliable and unambiguous evidence (e.g. a woman infers that a wolf is outside, based on seeing large, fresh paw-prints, in good light). In **Seeing Weak Evidence** (SW) stories, the protagonist makes the same inference based on indistinct and unreliable visual evidence (e.g. the woman sees some old markings in the dirt in dim light). In **Hearing First-person Evidence** (HF) stories, the protagonist draws her conclusion based on strong aural evidence (e.g. the woman hears a distinctive growl on a quiet afternoon). Finally, in **Hearsay Evidence** (HS) stories, the protagonist comes to her conclusion based on someone else's report of evidence (e.g. the woman is told by her friend that there are large, fresh paw prints in the dirt). Individual participants heard each target story in only one of the four epistemic conditions, forming a matched and counterbalanced design; every target story occurred in all four conditions across participants.

In half of the stories in each condition, the protagonist felt a **Negative** emotion based on her inference (e.g. that there is wolf outside), whereas in the other half of the stories, the protagonist would feel a **Positive** emotion based on her inference (e.g. that she is getting a puppy for her birthday, Fig. 1). After each story, participants heard a recording of a male speaker ask "Happy or sad?" Participants indicated whether the main character in the story felt happy or sad, using a button press (left/right). During the response period, a screen with a happy (left) and sad (right) face appeared, to remind participants which button to press. Reaction time was measured from the onset of the question "Happy or sad?"

Stories were presented in a pseudorandom order, where condition order was counterbalanced across runs and subjects, and no condition was immediately repeated. During the stories, one of eight task-irrelevant abstract color-patterns was randomly displayed. Rest blocks of 12 s occurred four times in each run. The total experiment of five runs (6.2 min each; 10 stories per run) lasted 31 min.

2.3.2. Key comparisons

Independent behavioral ratings (see SM and Table S1) verified that the intended manipulations were effective: Seeing Strong Evidence stories were rated as having higher justification than Seeing Weak Evidence ($t(77.2) = 7.15$, $p < 0.001$, $d = 1.6$). In contrast, there was no difference in the belief justification between hearing first-person (HF) and hearsay (HS, $t(77.2) = 0.94$, $p = 0.35$), nor between seeing and hearing ($t(149.4) = -1.05$, $p = 0.29$). Note that there was a reliable difference in the justification between the Seeing Strong-Evidence stories and the First-person Hearing stories ($t(78) = 2.14$, $p = 0.04$, $d = 0.48$). In general sighted people treat visual evidence as stronger than auditory evidence. As a result, this comparison is confounded between modality and justification.

We therefore tested three features of epistemic evaluation of beliefs: **Justification** contrasts seeing strong evidence (SS) and seeing weak evidence (SW), holding fixed the modality, directness, valence, and conclusion. **Modality** contrasts whether the protagonist received the information visually (SS + SW) or aurally (HF + HS), holding fixed the conclusion and the justification of the belief. **Directness** contrasts whether the protagonist heard the evidence him/herself (HF), vs. was told about the evidence (HS), holding fixed the evidence and the conclusion, justification and source modality. We hypothesized that these epistemic features might be represented in bilateral TPJ.

The **Valence** of the protagonist's emotion was the difference between the positive and negative stories. The protagonists of positive stories were rated as significantly happier than the protagonists of negative stories (Positive: 6.13 ± 0.06; Negative: 1.69 ± 0.03; $t(110.5) = 63.35$, $p < 0.001$, $d = 10.02$ (large)). By contrast, the four epistemic conditions were matched on emotional valence ($F(3,156) = 0.03$, $p = 0.99$), and across all stories and conditions (SS, SW, HF, HS), positive stories and negative stories had similar evidence justification (Positive: 4.4 ± 0.1; Negative 4.7 ± 0.1; $t(110.5) = 1.6$, $p = 0.11$). We hypothesized that valence would be represented in MPFC. Previous work has provided evidence that dorsal parts of MPFC represent valence (Peelen et al., 2010; Skerry and Saxe, 2014, 2015), and that affective theory of mind reasoning depends on VMPFC (Amodio and Frith, 2006; Etkin et al., 2011; Hynes et al., 2006; Sebastian et al., 2012; Shamay-Tsoory et al., 2006; Shamay-Tsoory and Aharon-Peretz, 2007; Leopold et al., 2011; Shamay-Tsoory, 2011); thus, we tested for representations of valence in dorsal, middle, and ventral medial prefrontal cortex.

2.3.3. Experiment 2: replication and generalization

To test the replicability of the evidence justification results of Experiment 1, and their robustness to variation in the stimulus and task, we analyzed an existing dataset that contained a conceptually similar manipulation. Young et al. (2010) asked participants to make moral judgments of a protagonist's actions (e.g. "Grace puts the powder in her friend's coffee") based on justified or unjustified beliefs (Young et al., 2010). In this study, participants read 54 stories in the scanner. Each story was broken into three sections: (1) the background, which set the stage (identical across conditions; 6 s), (2) the protagonist's belief (identical across conditions), and the justification for the belief (**weak**, **strong,** or unspecified; 4 s), and (3) the conclusion, which detailed the protagonist's action and the final outcome (bad: the protagonist's action led to physical injury or death of another character; 1/3 stories; or neutral: no harm to anyone; 2/3 of the stories; 6 s; Fig. 3). Only stories describing strong or weak evidence are analyzed here, collapsing across outcome. All stories are available for download (http://saxelab.mit.edu/stimuli.php) and in SM.

Belief justification was manipulated by describing the belief as based on strong evidence (e.g. "Grace thinks the white powder is sugar, because the container is labeled 'sugar'") or based on insufficient evidence (e.g. Grace thinks the white powder is sugar, though there's no label on the container."). The original publication of these data examined the magnitude of the response in theory of mind brain regions (Young et al., 2010). In the current analysis, we tested whether the *pattern* of response in any ToM region (particularly the RTPJ and RMSTS, given the results of Exp. 1) distinguished between others' beliefs that were based on strong vs. insufficient evidence (Fig. 3).

Each story appeared once in each condition, forming a matched and counterbalanced design. Individual participants saw each story in only one of the 9 conditions; every story occurred in all conditions, across participants. Stories were presented in a pseudorandom order; conditions were counterbalanced across runs and participants, and 14 s rest blocks were interleaved between stories. After each story, participants responded to the question "How morally blameworthy is [the agent] for [performing the action]?" on a 4-point scale (1-not at all, 4-very much), using a button press. Stories were presented in white, 24-point font on a black background. Nine stories were presented in each 5.1-min run, with six runs total (30.6 min).

2.3.4. Theory of mind localizer

Participants in both experiments were also scanned on a localizer task designed for identifying theory of mind brain regions in individual participants. Participants read verbal narratives about thoughts ('Belief'), vs. about physical representations like photographs and maps ('Photo' (Dodell-Feder et al., 2011); see SM).

2.4. Acquisition and preprocessing

FMRI data were collected in a 3 T Siemens scanner at the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT, using a 32-channel phased array head coil (Exp. 1) and a 12-channel head coil (Exp. 2).

We collected a high-resolution (1 mm isotropic) T-1 weighted MPRAGE anatomical scan, followed by functional images acquired with a gradient-echo EPI sequence sensitive to blood-oxygen-level-dependent (BOLD) contrast (repetition time (TR) = 2s, echo time (TE) = 40 ms, flip angle = 90°, voxel size 3 × 3 × 3mm, matrix 64 × 64, 26 near-axial slices). Slices were aligned with the anterior/posterior commissure and provided whole-brain coverage (excluding the cerebellum). For steady state magnetization, the first 4 s of each run were excluded.

Data were motion corrected, realigned, normalized onto a common brain space (rigid rotation and translation, MNI template, SPM8), spatially smoothed using a Gaussian filter (FWHM 5 mm) and subjected to a high-pass filter (128 Hz). See SM for motion and artifact analysis.

2.5. fMRI analysis

2.5.1. Defining individual subject ROIs for theory of mind

Individual ROIs were picked using the data generated by the ToM localizer (Dodell-Feder et al., 2011). Individually-tailored functional regions of interest (ROIs) were defined in right and left temporo-parietal junction (R/LTPJ), right middle superior and bilateral anterior temporal sulcus (RMSTS, R/LASTS), medial precuneus (PC), and dorsal-, middle-, and ventro-medial prefrontal cortex (D/M/VMPFC) (9 regions total). ROIs were defined based on in the first-level contrast image, as all voxels within a 9 mm radius of the peak voxel that passed threshold (Belief > Photo, $p < 0.001$ uncorrected, $k > 20$; SM and Fig. S1).

2.5.2. Defining language ROIs

In order to test for regional specificity of our results to ToM regions, we tested our four key comparisons (justification, modality, directness, and valence) in 14 language regions, which span much of language-responsive cortex. These group-level ROIs were based on the parcels from Fedorenko et al. (2010), which defined functional regions in which activity in a sentence > non-words contrast was found most consistently across subjects (Fedorenko et al., 2010) (see SM and Fig. S1). Language regions are a relevant set of control regions, because they are likely recruited to process the aural verbal stimuli, and represent abstract features and semantic content of the stories.

2.5.3. Within-ROI multivariate analysis (MVPA)

We conducted within-ROI multi-voxel pattern analysis (MVPA). To reduce the dimensionality of the number of features (voxels) relative to the number of items, feature selection was used to find voxels with each ROI most likely to contain task-related signal (Pereira et al., 2009). Using data from all runs of the main experiment, an unbiased task vs. rest ANOVA identified up to 100 most active voxels in each ROI, in which the response to all conditions differed most reliably from rest (ranked by F-statistic) (Mitchell et al., 2004). Because all conditions are modeled together, this selection criterion did not bias the outcome of classification between conditions ("peeking") and allowed us to use the same voxels across cross-validation folds (see SM for further details).

Response patterns in each ROI were classified using a linear support vector machine (SVM), which plausibly models the readout of neural populations (Butts et al., 2007; DiCarlo and Cox, 2007; Naselaris et al., 2011). For each trial, we calculated the average BOLD signal in each voxel in the ROI (high-pass filtered (within run; 128 Hz), linearly detrended (across runs), z-scored per voxel; Fig. S2a), measured during the "evidence" portion of the story. For Exp 1, the "evidence" began 17s into the story, and lasted 4–16s; in Exp 2, the "evidence" began 7s into the story and lasted 4s; in both cases, responses were averaged starting 4s later to account for hemodynamic lag. This procedure resulted in a temporally-compressed neural "pattern" for each trial: a vector of BOLD responses across voxels in the ROI (Fig. S2b).

For each classification, we used all trials in all-but-one runs to train the model, and all trials in the left-out run to test classification (Table S2). This procedure was iterated. We then averaged across folds to produce one classification rate, per ROI and per participant. Statistics were calculated across participants: a region could reliably distinguish a feature if binary classification accuracies were significantly above chance, across participants (0.50; one-sample, one-tailed *t*-test; see SM and Table S2 for further details).

2.5.4. Correcting for multiple comparisons and determining specificity of effects

In Exp. 1, we initially tested for significant decoding of source modality and justification in bilateral TPJ (two regions/tests, Bonferroni corrected α-level = 0.025), and valence in D/M/VMPFC (three region/tests, α-level = 0.017). In Exp. 2, we tested the replicability and robustness of the justification results in Exp. 1. In both experiments, we subsequently tested the remaining ToM regions and the 14 language

regions, in order to determine specificity of the results to the a priori hypothesized ROIs.

Finally, in both experiments, we tested whether there was any reliable difference between regions in the information they contained. In each subject, all 23 regions were ranked by their classification accuracy for each dimension; we used a Wilcoxon Rank Sum test to determine whether the regions that showed successful classification reliably contained more information than other regions. This measure allowed us to compare across regions without the explosion of multiple comparisons created by pairwise tests among 23 regions for 4 dimensions, and is more tolerant of missing data (i.e. participants in whom an ROI was not identified) than a within-subject analysis of variance.

2.5.5. Item-wise classification: testing for continuous representations

In addition to looking at subject-wise classification accuracies, we also calculated item-wise classification. For each item, classification of justification is defined as the proportion of times an item was classified as "strong evidence", across participants; modality classification is defined as the proportion of times an item was classified as visual evidence. We asked whether ROIs represent continuous information about the distinction (e.g. very strong evidence vs. somewhat ambiguous evidence), and whether item-wise classification scores (e.g. proportion of times an item was classified as "strong") were predicted by item-wise behavioral ratings (1–7 scale of "how good is the evidence"?). See SM for further details.

3. Results

3.1. fMRI task: classification results

Do regions implicated in reasoning about other minds contain explicit, abstract representations of features of others' belief formation process? We test if bilateral TPJ represents belief justification and source modality, and if MPFC represents valence. To test for specificity of significant results, we test the remaining ToM regions and 14 control regions implicated in language processing (Fedorenko et al., 2010).

3.1.1. Justification

The key open question that this study addresses is whether bilateral TPJ tracks information directly related to epistemic reasoning. From the pattern of neural responses across voxels in individually defined ROIs, we successfully classified stories describing justified versus unjustified beliefs in RTPJ (accuracy = 0.56(0.03), $t(16) = 2.4$, $p < 0.014$; α-level = 0.025 for two regions/tests).

Next, we test for the specificity of this result by testing the remaining ToM regions and the 14 language regions. We find a significant effect in RMSTS (0.58(0.04), $t(16) = 2.2$, $p = 0.023$, Fig. 2b), and no significant results in the other regions tested (all $p > 0.1$). Comparing across all regions, there was reliably more information about justification in the neural patterns in RTPJ (Wilcoxon Sum Rank test, $W = 107$, $p = 0.023$) and RMSTS ($W = 97$, $p = 0.019$) than in the other 21 regions.

Is this information categorical or continuous? Independent behavioral ratings, collected via Amazon's Mechanical Turk (Buhrmester et al., 2011) (see SM) of the justification of the belief showed that single items varied in the extent to which the evidence was considered strong or weak. Behavioral ratings predicted how likely each individual story was to be classified as depicting strong evidence in both RTPJ ($r(78) = 0.29$, $p = 0.009$) and RMSTS ($r(78) = 0.28$, $p = 0.01$, Fig. 2e). Continuous behavioral ratings explained significant variance in the item-wise neural classification accuracies, even after accounting for the binary condition labels (RTPJ: Condition: $\beta = 0.16 \pm 0.1$, $t = 1.7$, $p = 0.11$; Behavioral rating: $\beta = 0.22 \pm 0.1$, $t = 2.3$, $p = 0.038$; RMSTS: Condition: $\beta = 0.2 \pm 0.08$, $t = 2.2$, $p = 0.035$; Behavioral rating: $\beta = 0.17 \pm 0.08$, $t = 2.1$, $p = 0.048$).

3.1.2. Modality

Replicating a prior study (Koster-Hale et al., 2014), we found distinct patterns of response to beliefs based on visual vs. auditory evidence in LTPJ (0.59(0.03), $t(16) = 2.8$, $p = 0.007$); and a marginal effect in RTPJ (accuracy = 0.55(0.03), $t(16) = 2.1$, $p = 0.027$) after correcting for 2 regions/tests (α-level = 0.025, Fig. 2b); and in no other region (all $p > 0.15$). Across all 23 regions, more information about modality was consistently found in LTPJ (Wilcoxon Sum Rank test, $W = 124$, $p = 0.002$) and RTPJ ($W = 90$, $p = 0.047$).

Are representations of evidence modality in RTPJ and LTPJ independent? Classification scores for modality in RTPJ and LTPJ were correlated across items ($r(157) = 0.21$, $p = 0.009$) and across participants ($r(15) = 0.51$, $p = 0.04$), suggesting related representations of modality across regions. However, lexical concreteness scores (which were overall higher for visual evidence than auditory evidence) predicted item-wise classification accuracy for modality in LTPJ ($r(158) = 0.18$, $p = 0.02$) but not RTPJ ($r(158) = -0.004$, $p = 0.96$; difference: $t = 1.9$, $p = 0.03$), suggesting that classification in LTPJ may be partially driven by a representation of the concreteness of the evidence in the story, rather than the sensory modality alone.

3.1.3. Justification vs. modality in RTPJ

The evidence here suggests that RTPJ represents both belief justification and evidence modality (Koster-Hale et al., 2014). These two dimensions are often correlated in real-world situations (visual evidence tends to be perceived as stronger), although they were manipulated orthogonally in these stimuli. One possibility is that the RTPJ encodes only one of these two dimensions, and the other dimension is distinguished by proxy. We tested this hypothesis in three ways. First, we asked whether an item that was rated as more *justified* was more likely to be classified as *visual* based on the pattern of response in the RTPJ. We found no such correlation, either in all 160 items, or within the 80 visual items alone (all items: $r(158) = 0.04$, $p = 0.6$; visual items: $r(78) = -0.01$, $p = 0.9$). Second, there was no correlation, across participants, between the classification accuracies in RTPJ for modality vs. justification ($r(15) = -0.22$, $p = 0.4$). Finally, behavioral ratings of evidence quality were better predictors of neural classification of justification than of modality (difference of correlations, $z(78) = 1.93$, $p = 0.049$). Together, these results suggest that RTPJ contains independent and orthogonal representations of the justification and source modality of others' beliefs.

3.1.4. Directness

Do any regions distinguish between others' beliefs based on first-person auditory evidence vs. hearsay (another person's report), independent of justification and modality? We could not classify stories into these categories in any region (all $p > 0.14$), and no region showed reliably more information about directness than any other region (all $p > 0.13$).

3.1.5. Valence

The pattern of neural response in the VMPFC reliably classified stories resulting in positive vs. negative emotions (accuracy = 0.57(0.04), $t(12) = 1.9$, $p = 0.038$). This effect is marginal when correcting for multiple comparisons ($\alpha = 0.017$, for 3 regions (D/M/VMPFC)/tests), but is consistent with previous studies. Information about valence was more present in VMPFC than in the other 22 regions ($W = 77$, $p = 0.013$). Again, we asked whether these representations were categorical or continuous. Independent behavioral ratings of the valence of the protagonist's emotion in each story predicted how likely that story was be to be classified as positive valence by the VMPFC ($r(158) = 0.21$, $p = 0.009$, Fig. 2e), even after accounting for the binary condition labels (Condition: $\beta = 0.13 \pm 0.07$, $t = 2.2$, $p = 0.047$; Behavioral rating: $\beta = 0.15 \pm 0.07$, $t = 2.2$, $p = 0.02$).

3.1.6. Functional dissociation: RTPJ vs. VMPFC

Does RTPJ contain reliably more information about evidence quality

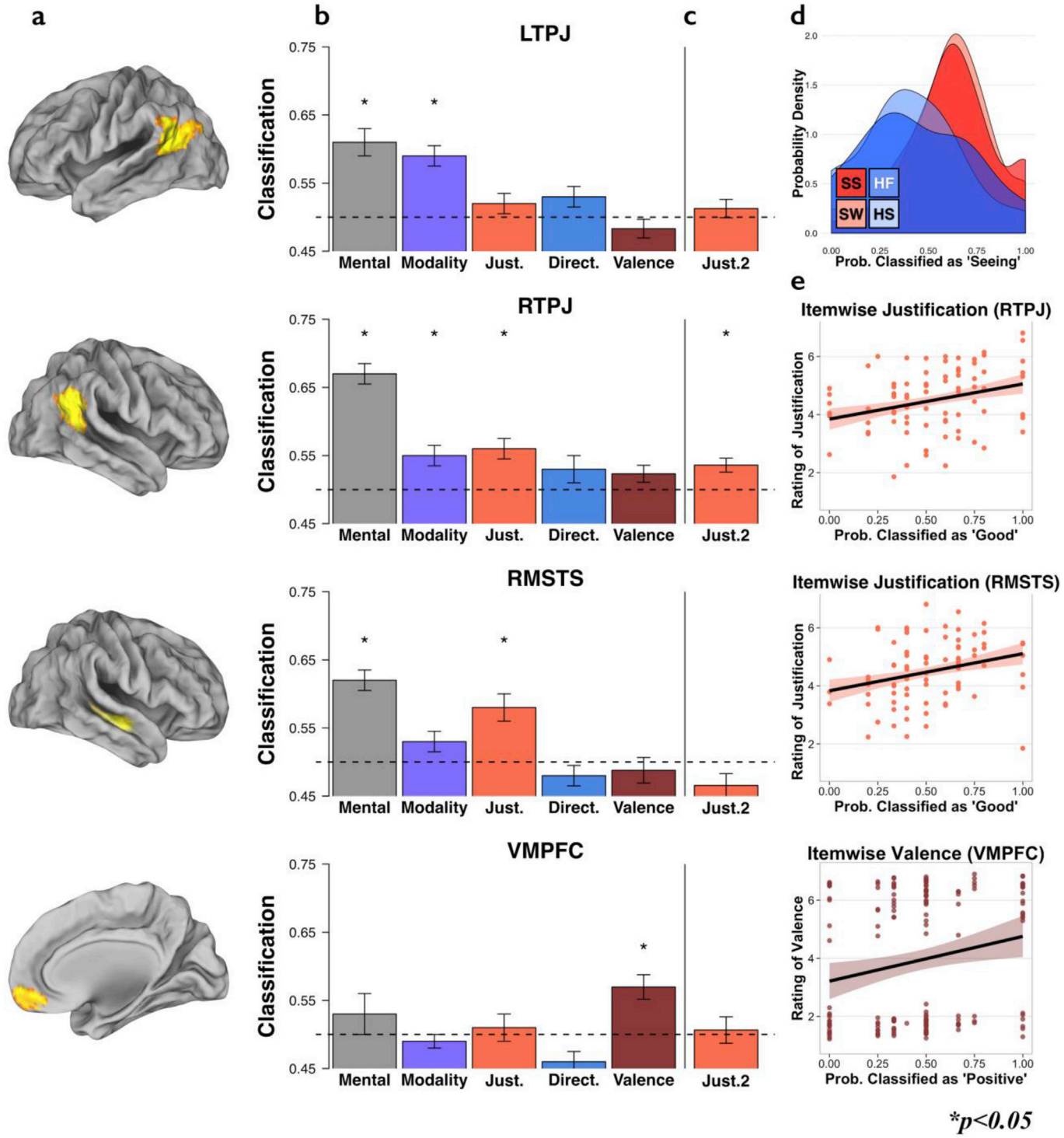

Fig. 2. **(a)** ROIs **(b)** Classification accuracy in Exp. 1 for **Mental**: mental vs. physical stories (see SM), **Modality:** visual vs. auditory modality (S,H), **Justification:** strong vs. weak evidence (SS, SW), **Directness:** first person vs. hearsay (HF, HS), and **Valence:** positive vs. negative emotion, **(c)** Classification accuracy in the replication study for **Justification:** strong vs. weak evidence. **(d)** LTPJ: density plot of classification accuracies by condition. Modality is not a continuous feature and thus not compared to behavior; rather we observe striking similarity in the classification accuracy in **LTPJ** between seeing conditions (red: SS and SW) and hearing conditions (purple: HF HS). **(e)** RTPJ, RSTS, VMPFC: Correlation between item-wise classification scores and behavioral ratings. In **RTPJ** and **RMSTS** there were significant correlations between justification classification scores (how many times an item was scored as "strong", across participants) and behavioral judgments (how good is the character's evidence?). In **VMPFC**, there was a significant correlation between valence classification score (how many times an item was scores as "positive") and behavioral judgments (how happy does the character feel?). (***p* < *0.05***).

than VMPFC, and does VMPFC contain reliably more information about valence than RTPJ? Converging with the results from the Wilcoxon sum tests, item-wise valence scores were significantly more related to classification accuracies in the VMPFC than in the RTPJ (difference of correlations, $t = 2.4$, $p = 0.02$). Conversely, item-wise classification by justification was significantly more related to ratings of justification in RTPJ than in VMPFC (difference $t = 2.15$, $p = 0.02$), confirming that these regions contain information about distinct aspects of the stories.

3.1.7. Shared representations across regions

When two regions could successfully classify the same feature, we tested whether those representations were independent or redundant. Patterns of activation in both RMSTS and RTPJ discriminated the justification of belief, but there was no correlation between the classification accuracies in RTPJ and RMSTS when collapsed by item ($r(78) = 0.18$, $p = 0.12$) or by subject ($r(15) = -0.19$, $p = 0.48$). The classification accuracy of RTPJ and RMSTS made independent contributions towards explaining the behavioral ratings (RTPJ: $\beta = 1.0 \pm 0.4$, $t = 2.3$, $p = 0.02$; RMSTS: $\beta = 1.1 \pm 0.5$, $t = 2.2$, $p = 0.03$), suggesting that RTPJ and RMSTS encode different aspects of belief justification. By contrast, both RTPJ and LTPJ could classify modality of evidence, and these regions' classification scores were correlated across items ($r(157) = 0.21$, $p = 0.009$) and across participants ($r(15) = 0.51$, $p = 0.04$), suggesting that RTPJ and LTPJ contain related representations of source modality.

3.1.8. Summary of results in language control regions

None of the control regions showed successful classification of any of the tested distinctions (all $T < 1.4$, $p > 0.05$ uncorrected), consistent with the hypothesis that this information is specific to ToM regions.

3.1.9. Effects of age and gender

None of the reported results differ by age (all $rs < 0.51$, all $ps > 0.07$) or gender (all $ts < 1.4$, all $ps > 0.19$).

3.2. Replication: generalization to another manipulation of justification

A key novel result of Exp. 1 is the evidence of a representation of the justification of others' beliefs in RTPJ and RMSTS. To test the replicability of this result, and its robustness to variation in the stimulus and task, we reanalyzed published data that contained a similar manipulation (Young et al., 2010). In this conceptual replication, the "bad evidence" consisted of missing or misleading evidence (Fig. 3), in contrast to Exp. 1, which manipulated the protagonist's perceptual access and the extent to which the evidence unambiguously supported the conclusion. In addition, the stimulus manipulation in Exp. 2 affected 2–4 *words* in each story, so that the difference between conditions was a minimal pair (e.g. "**Because** the container is labeled **sugar**" vs. "**Although** there is **no** label on the container", see Fig. 3). Thus, we can use these data to ask not only whether we find a reliable difference in neural patterns for justified and unjustified beliefs in a new set of participants, with new stimuli and a new task, but also whether this representation extends to a distinct and minimal manipulation of justification.

As in Exp. 1, justified versus unjustified beliefs were successfully classified based on the neural pattern of response in RTPJ (accuracy = 0.54(0.02), $t(17) = 1.8$, $p < 0.05$, Fig. 2c), but not RMSTS (accuracy = 0.47(0.03), $t(15) = -0.99$, $p = 0.83$; Wilcoxon signed-rank $W = 6$, $p = 0.70$; because the distribution of the RMSTS differed significantly from a normal distribution (Lilliefors test for normality, $K = 0.26$, criterion = 0.21, $p = 0.004$), we report results from the relevant parametric and non-parametric tests). RMSTS carried reliably less information about justification than RTPJ (Wilcoxon signed-rank $W = 86.5$, $p = 0.04$). Moreover, RMSTS carried significantly less information than it did in Exp. 1 (Wilcoxon signed-rank $W = 75.5$, p-value = 0.03).

In order to test for the specificity of this effect in the RTPJ, we additionally tested the other ToM ROIs and the 14 control ROIs: no other ROI showed this distinction (all classification accuracies <51%, p>0.3).

4. Discussion

Humans have a rich and detailed model of other minds, which includes understanding both what actions people will take given their beliefs and desires, and what beliefs people will form given their environment. We find evidence that, mirroring the processes observed in sensory systems (DiCarlo et al., 2012; Kamitani and Tong, 2005; Kourtzi, 2001; Lafer-Sousa and Conway, 2013; Tanaka, 1993), social stimuli are represented in distinct and distributed formats across the human brain. Specifically, we observed a spatial and functional dissociation epistemic features of another person's beliefs, represented in RTPJ, and the valence of their beliefs, represented in VMPFC. These results converge with one hypothesized architecture of theory of mind, which distinguishes between epistemic and motivational components of social reasoning (Schlaffke et al., 2014; Schnell et al., 2011) (Amodio and Frith, 2006; Etkin et al., 2011; Hynes et al., 2006; Sebastian et al., 2012; Shamay-Tsoory et al., 2006; Shamay-Tsoory and Aharon-Peretz, 2007), is complementary with the division in computational models between belief formation and value-based planning (Baker et al., 2017, 2009; Bello, 2012; Jara-Ettinger et al., 2012), and provides an overarching framework for interpreting past neuroimaging results.

The key contribution of the current study is to characterize the features and representations of epistemic ToM, and dissociate those representations from representations of motivation and valence. Prior studies have reported that MPFC contains information about the valence of another person's experience (Chavez and Heatherton, 2015; Chib et al., 2009; Chikazoe et al., 2014; Kable and Glimcher, 2007; Peelen et al., 2010; Skerry and Saxe, 2014; Winecoff et al., 2013); we replicate that finding here. We extend this work with evidence that neural populations in VMPFC represent valence as a continuous dimension, ranging from positive to negative, and correlated with behavioral ratings.

Critically, in Exp. 1, we find that these motivational representations are distinct from epistemic representations of the same stimuli. We tested for two types of epistemic representation: the modality of another person's evidence (i.e. whether the other person saw or heard evidence (Koster-Hale et al., 2014)), and the justification of their belief (whether the evidence was strong or weak support for the conclusion they made).

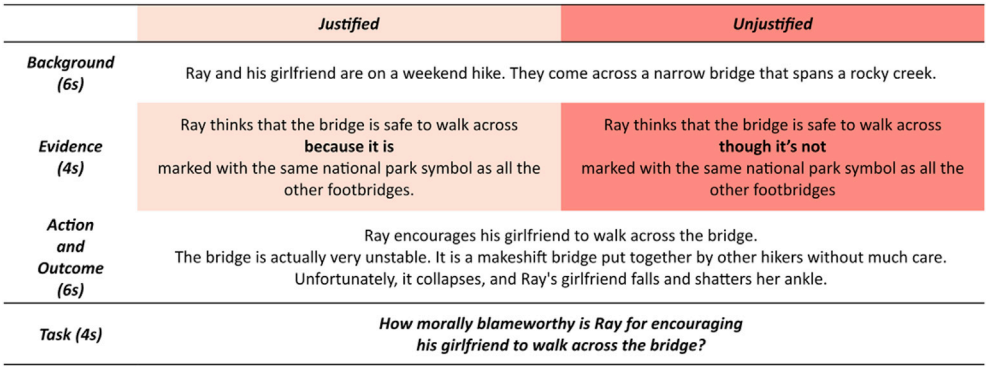

Fig. 3. Example stimuli from Exp. 2. Here, beliefs were unjustified not because of obscured and ambiguous evidence as in Exp. 1, but because of missing or misleading evidence (see SM). After each story, participants responded to the question "How morally blameworthy is [the agent] for [performing the action]?" on a 4-point scale (1-not at all, 4-very much), using a button press.

Both of these distinctions could be decoded in RTPJ but not MPFC.

RTPJ (in Exp. 1 and 2) and RMSTS (in Exp. 1 only) contain information about the justification of a person's belief. These patterns reflect spontaneous evaluation of the character's belief formation process; participants' explicit task focused attention on other aspects of the stories (emotion in Exp. 1, moral blame in Exp. 2). In both regions, belief justification was represented as a continuous, not a binary, feature and was correlated with independent behavioral ratings of each story. Intriguingly, the RTPJ and RMSTS contain at least partially distinct and complementary information about belief justification, hinting at a finer grain of distinctions between mental states. An alternative possibility is that the classification of justification in RMSTS in Exp. 1 was spurious, and therefore did not replicate in Exp. 2.

People often conflate modality with justification: visual evidence may seem to be stronger evidence than perceptions in other modalities ("seeing is believing", "my eyes don't deceive me"). By contrast, modality and justification involved distinct neural representations: LTPJ contained information only about modality, and within RTPJ, the representations of modality and justification were independent. The distinction between modality and justification within RTPJ in particular suggests that a multidimensional feature space of epistemic evaluation may be implemented in the RTPJ.

Similarly, directness of evidence and justification are also often conflated, especially when the source of hearsay evidence is perceived to be unreliable (Kovera et al., 1991; Miene et al., 1993; Olson, 2003). In the current experiment, the first-person and hearsay conditions were matched on evidence justification, suggesting that the informants in the hearsay condition were overall perceived to be reliable. This design enabled us to test for representations of directness per se, controlling for justification; we did not find evidence that ToM brain regions represent the directness of evidence. Future work that manipulates justification and directness orthogonally (e.g. in a 2 × 2 design), or manipulates different aspects of directness (e.g. is the informant reliable? What is the relationship between the protagonist and the informant?) will be informative for understanding the neural representation of direct evidence versus hearsay.

The representational dimensions revealed by these analyses are highly abstract, generalized representations of a shared latent feature across otherwise unique and heterogeneous verbal stories. All classification analyses involved training and testing on unique stimuli (not repeated instances of the same stimulus). Because each story occurred in all conditions (across subjects), differences between conditions were minimal, limited to the words describing the character's belief formation. Those diagnostic words were variable within a condition, across stories (e.g. two examples of unjustified beliefs: "*Bella tried to peer through a crack in the door. In the very* dim *light, Bella squinted to see his eyes close.*"; "*The classroom was large and crowded. Across the room, Dillon was pointing at something.*"). The manipulation in Exp. 2 was even more minimal: changes of just a few words in each story (e.g. justified: "*because it is marked*" vs. unjustified: "*although it is not marked*"). Thus, across different participants, stimuli, and tasks, patterns of neural response spontaneously distinguish at least one abstract feature of the character's belief formation process: justification of evidence. Taken together, these studies provide evidence of an abstract representation of others' belief formation process complementary to reasoning about motivation and values.

The contrast between epistemic features represented in TPJ and motivational features represented in VMPFC (Exp. 1) helps to clarify another puzzle in the literature, regarding the representation of accidental harm. One set of experiments described a protagonist as causing harm knowingly versus unknowingly (e.g. "you *absolutely knew/had no idea* about your cousin's allergy when you served him the peanuts). Information about this distinction was found in RTPJ, and predicted participants' moral judgments of the protagonist (Koster-Hale et al., 2013). By contrast, in a second experiment, when an action was depicted as on purpose versus by accident (e.g. deliberately pushing someone versus tripping and falling against them), activation was different in the VMPFC. Developmental change in VMPFC was associated with developmental reductions in blame for the accidents (Decety et al., 2012). These two sets of results are compatible when viewed in light of the proposed representational architecture for ToM: RTPJ contains information about what the protagonist knew or should have known, before acting intentionally (i.e. an aspect of their belief formation); whereas the VMPFC is sensitive to whether the action was consistent with the protagonist's goals (i.e. an aspect of their value-based planning).

One prominent use of MVPA has been to find information about a stimulus or task outside of the regions that show peak or selective responses (O'toole et al., 2005). By contrast, we found that both motivational and epistemic features of the stimuli were represented in brain regions associated with ToM, but not in brain regions associated with language processing. These results converge with multiple prior reports that information relevant to distinctions within ToM appears to be preferentially represented in the same regions that show a robust response to ToM overall (Koster-Hale et al., 2014, 2013; Skerry and Saxe, 2015, 2014; Tamir et al., 2016). Mounting recent evidence suggests a distinct partition of the TPJ is recruited for mental state reasoning; given the use of a validated functional localizer, these results pertain to this partition (Igelstro m et al., 2016; Mars et al., 2012). This convergence between MVPA and evoked responses suggests that knowledge of other minds may be implemented in representational spaces distinct from other aspects of conceptual and linguistic processing.

However, null results in MVPA must always be interpreted with caution (e.g. due to limits in spatial scale (Dubois et al., 2015)). Because each fMRI voxel contains thousands of neurons, MVPA can only detect relatively low-frequency spatial patterns of neural responses, and many distinct neural populations are intermingled beyond this resolution (Freeman et al., 2011; Op de Beeck, 2010). Furthermore, evoking rich and specific mental states requires relatively long and complex stimuli. We classified average neural responses to a whole sentence, presented in the context of a longer narrative, and thus combined across many cognitive processes. As a result, classification results must be interpreted as a lower bound on the information available in each region (Kriegeskorte and Kievit, 2013).

Moreover, finding evidence of a series of distinctions is merely a small step into explaining the full variance of these regions. Discovering the complete representational spaces that structure our knowledge of other minds poses a major challenge. Observers' inferences about others' mental states are flexible, generative, and fine-grained. Any individual experiment can only test an extremely sparse sample of possible hypothesized representations. For example, in addition to the features described above, prior research has shown that the RTPJ contains information about a character's history of cooperation (Behrens et al., 2008) and likely future action (Carter et al., 2012).

As one approach to that challenge, two recent studies have used data-driven feature-discovery methods to characterize representational spaces in the ToM network. Tamir and colleagues found a single dimension that captures substantial variance in the pattern of local responses in all regions of the ToM network to 60 distinct states (attributed to an unnamed target) (Tamir et al., 2016). The authors call this dimension "social impact": it ranges from highly social, high arousal states like playfulness, lust, dominance, and embarrassment at one end, to solitary, low-arousal states like exhaustion, laziness, self-pity, and relaxation at the other. Two additional (and largely orthogonal) dimensions also explained variance in the pattern of responses in the ToM network: valence (satisfaction and inspiration, versus disarray and distrust), and "rationality" (planning and decision, versus ecstasy and disgust). Though Tamir and colleagues did not test whether these distinct representational spaces within ToM are associated with distinct neural loci, this study provides converging evidence that valence and rationality are distinct aspects of our intuitive concepts of other minds.

Using a similar approach, Skerry and Saxe found that patterns of response in the ToM network, including RTPJ and MPFC, could classify 200 unique short verbal narratives into classes described by twenty

distinct emotion labels (e.g. furious, jealous, grateful, proud) (Skerry and Saxe, 2015). The similarity space of neural patterns in this network was best captured by abstract features of the situation, and was not reducible to more primitive affective dimensions such as valence and arousal. Features that explained significant variance in the neural response included whether the event influenced the protagonist's significant relationships, would be repeated in the future, affected the protagonist's life in the long run, and/or was caused by the protagonist or by other people. These features are related, but not reducible, to Tamir and colleague's concept of social impact. In addition, Skerry and Saxe found suggestive evidence of partially complementary (i.e. non-redundant) information represented in different regions within the network. As in Tamir et al., however, the stories in Skerry and Saxe did not manipulate the character's belief formation process, and so could not test distinctions within the epistemic component of ToM. Thus, future work should replicate and extend the dissociation we find evidence for here, between representations of epistemic features in TPJ and valence in VMPFC.

One challenge of the data-driven approach is that it is unlikely that mental state attributions are best described as simply a list of features; rather, they are likely representations with internal structure (Baker et al., 2017, 2009; Davidson, 1963), understood in terms of their computational role within a coherent explanatory theory (Carey, 2009; Gopnik and Wellman, 1994). Any representational similarity analysis operationalizes these representations as a "bag of features", more similar to the way concepts have been defined in prototype theory (i.e. graded categorization based on feature similarity to some category prototype or centroid (Rosch, 1973)). This representation lacks compositionality and cannot naturally encode logical or causal structure (Kording, 2014; Tervo et al., 2016). Even a simple propositional attitude (e.g. *The captain believes that he has found the sea monster*) is composed of an agent (the captain), an attitude (believes) and a propositional content (he has found the sea monster), and is causally connected to many other specific mental states (e.g. wanting to find the sea monster, feeling excited, deciding to set off in pursuit). Relatedly, inferences about beliefs necessarily depend on a rich body of world knowledge (e.g. sea monsters are dangerous but rare), so neural populations specific to theory of mind must interface with general-purpose semantic systems. It will be a challenge for future research to characterize the neural implementation of this highly complex system, and a combination of data-driven and hypothesis-driven approaches will be necessary to tackle this problem. We believe that, by providing evidence of functionally and spatially distinct components of theory of mind, this paper provides a crucial step in that process, and we look forward to future progress.

5. Conclusion

A cornerstone of the human capacity for social cognition is the ability to reason about the unobservable causal structure underlying other's actions: the person's intentions, beliefs, and goals. This study uses advanced functional neuroimaging techniques to probe neural computations underlying social reasoning. Building on existing computational models, we find that brain regions recruited during mental state reasoning contain neural signatures of epistemic and motivational components of theory of mind: the justification and source modality of the belief, and the valence of the resulting emotion. These representations reliably track behavior, and are encoded by distinct neural populations. By delineating these neural representations, we begin to probe the inner workings of mental state inference, and its distribution across the human brain.

Acknowledgments

We thank the Athinoula A. Martinos Imaging Center at the McGovern Institute for Brain Research at MIT, and the Saxe lab, especially Amy Skerry, Stefano Anzellotti, Dorit Kliemann, and Ben Deen for helpful discussion. We also gratefully acknowledge support of this project by NSF Graduate Research Fellowships (#0645960 to JKH, #1122374 to HR) and an NSF CAREER award (#095518), the National Institutes of Health (1R01 MH096914-01A1), and the Packard Foundation (Contract # 2008-333024 to RS).

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.neuroimage.2017.08.026.

References

Amodio, D.M., Frith, C.D., 2006. Meeting of minds: the medial frontal cortex and social cognition. Nat. Rev. Neurosci. 7, 268–277. http://dx.doi.org/10.1038/nrn1884.
Baker, C.L., Jara-Ettinger, J., Saxe, R., Tenenbaum, J.B., 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. Nat. Hum. Behav. 1, 0064.
Baker, C.L., Saxe, R., Tenenbaum, J.B., 2009. Action understanding as inverse planning. Cognition 113, 329–349.
Behrens, T.E.J., Hunt, L.T., Woolrich, M.W., Rushworth, M.F.S., 2008. Associative learning of social value. Nature 456, 245–249. http://dx.doi.org/10.1038/nature07538.
Bello, P., 2012. Cognitive foundations for a computational theory of mindreading. Adv. Cogn. Syst. 1.
Buhrmester, M., Kwang, T., Gosling, S.D., 2011. Amazon's mechanical Turk. Perspect. Psychol. Sci. 6, 3–5. http://dx.doi.org/10.1177/1745691610393980.
Butts, D.A., Weng, C., Jin, J., Yeh, C.-I., Lesica, N.A., Alonso, J.-M., Stanley, G.B., 2007. Temporal precision in the neural code and the timescales of natural vision. Nature 449, 92–95. http://dx.doi.org/10.1038/nature06105.
Carey, S., 2009. The Origin of Concepts. Oxford University Press.
Carrington, S.J., Bailey, A.J., 2009. Are there theory of mind regions in the brain? A review of the neuroimaging literature. Hum. Brain Mapp. 30, 2313–2335. http://dx.doi.org/10.1002/hbm.20671.
Carter, R.M., Bowling, D.L., Reeck, C., Huettel, S.A., 2012. A distinct role of the temporal-parietal junction in predicting socially guided decisions. Science 337, 109–111. http://dx.doi.org/10.1126/science.1219681.
Chavez, R.S., Heatherton, T.F., 2015. Multimodal frontostriatal connectivity underlies individual differences in self-esteem. Soc. Cogn. Affect. Neurosci. 10, 364–370. http://dx.doi.org/10.1093/scan/nsu063.
Chib, V.S., Rangel, A., Shimojo, S., O'Doherty, J.P., 2009. Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. J. Neurosci. 29, 12315–12320. http://dx.doi.org/10.1523/JNEUROSCI.2575-09.2009.
Chikazoe, J., Lee, D.H., Kriegeskorte, N., Anderson, A.K., 2014. Population coding of affect across stimuli, modalities and individuals. Nat. Neurosci. 17, 1114–1122. http://dx.doi.org/10.1038/nn.3749.
Clément, F., Koenig, M., Harris, P., 2004. The ontogenesis of trust. Mind Lang. 19, 360–379. http://dx.doi.org/10.1111/j.0268-1064.2004.00263.x.
Davidson, D., 1963. Actions, reasons, and causes. J. Philos. 60, 685. http://dx.doi.org/10.2307/2023177.
Decety, J., Michalska, K.J., Kinzler, K.D., 2012. The contribution of emotion and cognition to moral sensitivity: a neurodevelopmental study. Cereb. Cortex 22, 209–220.
DiCarlo, J.J., Cox, D.D., 2007. Untangling invariant object recognition. Trends Cogn. Sci. 11, 333–341. http://dx.doi.org/10.1016/j.tics.2007.06.010.
DiCarlo, J.J., Zoccolan, D., Rust, N.C., 2012. How does the brain solve visual object recognition? Neuron 73, 415–434. http://dx.doi.org/10.1016/j.neuron.2012.01.010.
Dodell-Feder, D., Koster-Hale, J., Bedny, M., Saxe, R., 2011. fMRI item analysis in a theory of mind task. NeuroImage 55, 705–712.
Dubois, J., de Berker, A.O., Tsao, D.Y., 2015. Single-unit recordings in the macaque face patch system reveal limitations of fMRI MVPA. J. Neurosci. 35, 2791–2802. http://dx.doi.org/10.1523/JNEUROSCI.4037-14.2015.
Etkin, A., Egner, T., Kalisch, R., 2011. Emotional processing in anterior cingulate and medial prefrontal cortex. Trends Cogn. Sci. 15, 85–93. http://dx.doi.org/10.1016/j.tics.2010.11.004.
Fedorenko, E., Hsieh, P.J., Nieto-Castanon, A., Whitfield-Gabrieli, S., Kanwisher, N., 2010. New method for fMRI investigations of language: defining ROIs functionally in individual subjects. J. Neurophysiol. 104, 1177–1194. http://dx.doi.org/10.1152/jn.00032.2010.
Freeman, J., Brouwer, G.J., Heeger, D.J., Merriam, E.P., 2011. Orientation decoding depends on maps, not columns. J. Neurosci. 31, 4792–4804. http://dx.doi.org/10.1523/JNEUROSCI.5160-10.2011.
Frith, C.D., Frith, U., 2012. Mechanisms of social cognition. Annu. Rev. Psychol. 63, 287–313. http://dx.doi.org/10.1146/annurev-psych-120710-100449.
Gopnik, A., Wellman, H.M., 1994. The theory. In: Presented at the an Earlier Version of This Chapter Was Presented at the Society for Research in Child Development Meeting, 1991. Cambridge University Press.
Harry, B., Williams, M.A., Davis, C., Kim, J., 2013. Emotional expressions evoke a differential response in the fusiform face area. Front. Hum. Neurosci. 7 http://dx.doi.org/10.3389/fnhum.2013.00692.

Hynes, C.A., Baird, A.A., Grafton, S.T., 2006. Differential role of the orbital frontal lobe in emotional versus cognitive perspective-taking. Neuropsychologia 44, 374–383. http://dx.doi.org/10.1016/j.neuropsychologia.2005.06.011.
Igelstro m, K.M., Webb, T.W., Kelly, Y.T., Graziano, M.S.A., 2016. Topographical organization of attentional, social, and memory processes in the human temporoparietal cortex. eNeuro 3. http://dx.doi.org/10.1523/ENEURO.0060-16.2016.
Jara-Ettinger, J., Baker, C.L., Tenenbaum, J.B., 2012. Learning what is where from social observations. In: Presented at the CogSci.
Kable, J.W., Glimcher, P.W., 2007. The neural correlates of subjective value during intertemporal choice. Nat. Neurosci. 10, 1625–1633. http://dx.doi.org/10.1038/nn2007.
Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 8, 679–685.
Koenig, M.A., Harris, P.L., 2005. The role of social cognition in early trust. Trends Cogn. Sci. 9, 457–459. http://dx.doi.org/10.1016/j.tics.2005.08.006.
Kording, K.P., 2014. Bayesian statistics: relevant for the brain? Curr. Opin. Neurobiol. 25, 130–133. http://dx.doi.org/10.1016/j.conb.2014.01.003.
Koster-Hale, J., Bedny, M., Saxe, R., 2014. Thinking about seeing: perceptual sources of knowledge are encoded in the theory of mind brain regions of sighted and blind adults. Cognition 133, 65–78. http://dx.doi.org/10.1016/j.cognition.2014.04.006.
Koster-Hale, J., Saxe, R., Dungan, J., Young, L.L., 2013. Decoding moral judgments from neural representations of intentions. Proc. Natl. Acad. Sci. 110, 5648–5653. http://dx.doi.org/10.1073/pnas.1207992110.
Kourtzi, Z., 2001. Representation of perceived object shape by the human lateral occipital complex. Science 293, 1506–1509. http://dx.doi.org/10.1126/science.1061133.
Kovera, M.B., Park, R.C., Penrod, S.D., 1991. Jurors' perceptions of eyewitness and hearsay evidence. Minn. L. Rev. 76, 703.
Kriegeskorte, N., Kievit, R.A., 2013. Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. 17, 401–412. http://dx.doi.org/10.1016/j.tics.2013.06.007.
Lafer-Sousa, R., Conway, B.R., 2013. Parallel, multi-stage processing of colors, faces and shapes in macaque inferior temporal cortex. Nat. Neurosci. 16, 1870–1878. http://dx.doi.org/10.1038/nn.3555.
Leopold, A., Krueger, F., dal Monte, O., Pardini, M., Pulaski, S.J., Solomon, J., Grafman, J., 2011. Damage to the left ventromedial prefrontal cortex impacts affective theory of mind. Soc. Cogn. Affect. Neurosci. 7, 871–880.
Lucas, A.J., Lewis, C., Pala, F.C., Wong, K., Berridge, D., 2013. Social-cognitive processes in preschoolers' selective trust: three cultures compared. Dev. Psychol. 49, 579–590. http://dx.doi.org/10.1037/a0029864.
Mars, R.B., Sallet, J., Schuffelgen, U., Jbabdi, S., Toni, I., Rushworth, M.F.S., 2012. Connectivity-based subdivisions of the human right "temporoparietal junction area": evidence for different areas participating in different cortical networks. Cereb. Cortex 22, 1894–1903. http://dx.doi.org/10.1093/cercor/bhr268.
Miene, P., Borgida, E., Park, R., 1993. The evaluation of hearsay evidence: a social psychological approach. Individ. group Decis. Mak. Curr. Issues 151–166.
Mitchell, T.M., Hutchinson, R., Niculescu, R.S., Pereira, F., Wang, X., Just, M., Newman, S., 2004. Learning to Decode Cognitive States from Brain Images. Mach. Learn. 57, 145–175. http://dx.doi.org/10.1023/B: MACH.0000035475.85309.1b.
Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. NeuroImage 56, 400–410. http://dx.doi.org/10.1016/j.neuroimage.2010.07.073.
O'toole, A.J., Jiang, F., Abdi, H., Haxby, J.V., 2005. Partially distributed representations of objects and faces in ventral temporal cortex. J. Cogn. Neurosci. 17, 580–590.
Olson, G., 2003. Reconsidering unreliability: fallible and untrustworthy narrators. Narrative 11, 93–109. http://dx.doi.org/10.1353/nar.2003.0001.
Op de Beeck, H.P., 2010. Probing the mysterious underpinnings of multi-voxel fMRI analyses. NeuroImage 50, 567–571. http://dx.doi.org/10.1016/j.neuroimage.2009.12.072.
Ozturk, O., Papafragou, A., 2016. The acquisition of evidentiality and source monitoring. Lang. Learn. Dev. 12, 199–230.
Papafragou, A., Li, P., 2001. Evidential morphology and theory of mind. In: Presented at the Proceedings from the 26th Annual Boston University Conference on Language Development. Cascadilla Press, Somerville, MA, pp. 510–520.
Papafragou, A., Li, P., Choi, Y., Han, C.-H., 2007. Evidentiality in language and cognition. Cognition 103, 253–299.
Peelen, M.V., Atkinson, A.P., Vuilleumier, P., 2010. Supramodal representations of perceived emotions in the human brain. J. Neurosci. 30, 10127–10134. http://dx.doi.org/10.1523/JNEUROSCI.2161-10.2010.
Pereira, F., Mitchell, T., Botvinick, M., 2009. Machine learning classifiers and fMRI: a tutorial overview. NeuroImage 45, S199–S209. http://dx.doi.org/10.1016/j.neuroimage.2008.11.007.
Robinson, E.J., Haigh, S.N., Nurmsoo, E., 2008. Children's working understanding of knowledge sources: confidence in knowledge gained from testimony. Cogn. Dev. 23, 105–118. http://dx.doi.org/10.1016/j.cogdev.2007.05.001.
Rosch, E.H., 1973. Natural categories. Cogn. Psychol. 4, 328–350. http://dx.doi.org/10.1016/0010-0285(73)90017-0.
Said, C.P., 2010. Graded representations of emotional expressions in the left superior temporal sulcus. Front. Syst. Neurosci. 4 http://dx.doi.org/10.3389/fnsys.2010.00006.
Said, C.P., Moore, C.D., Engell, A.D., Todorov, A., Haxby, J.V., 2010. Distributed representations of dynamic facial expressions in the superior temporal sulcus. J. Vis. 10 http://dx.doi.org/10.1167/10.5.11, 11–11.
Saxe, R., Powell, L.J., 2006. It's the thought that counts. Psychol. Sci. 17, 692–699. http://dx.doi.org/10.1111/j.1467-9280.2006.01768.x.
Schlaffke, L., Lissek, S., Lenz, M., Juckel, G., Schultz, T., Tegenthoff, M., Schmidt Wilcke, T., Brüne, M., 2014. Shared and nonshared neural networks of cognitive and affective theory-of-mind: a neuroimaging study using cartoon picture stories. Hum. Brain Mapp. 36, 29–39. http://dx.doi.org/10.1002/hbm.22610.
Schnell, K., Bluschke, S., Konradt, B., Walter, H., 2011. Functional relations of empathy and mentalizing: an fMRI study on the neural basis of cognitive empathy. NeuroImage 54, 1743–1754. http://dx.doi.org/10.1016/j.neuroimage.2010.08.024.
Sebastian, C.L., Fontaine, N.M.G., Bird, G., Blakemore, S.-J., De Brito, S.A., McCrory, E.J.P., Viding, E., 2012. Neural processing associated with cognitive and affective Theory of Mind in adolescents and adults. Soc. Cogn. Affect. Neurosci. 7, 53–63. http://dx.doi.org/10.1093/scan/nsr023.
Shamay-Tsoory, S.G., 2011. The neural bases for empathy. Neuroscience 17, 18–24.
Shamay-Tsoory, S.G., Aharon-Peretz, J., 2007. Dissociable prefrontal networks for cognitive and affective theory of mind: a lesion study. Neuropsychologia 45, 3054–3067. http://dx.doi.org/10.1016/j.neuropsychologia.2007.05.021.
Shamay-Tsoory, S.G., Tibi-Elhanany, Y., Aharon-Peretz, J., 2006. The ventromedial prefrontal cortex is involved in understanding affective but not cognitive theory of mind stories. Soc. Neurosci. 1, 149–166. http://dx.doi.org/10.1080/17470910600985589.
Skerry, A.E., Saxe, R., 2015. Neural representations of emotion are organized around abstract event features. Curr. Biol. 25, 1945–1954. http://dx.doi.org/10.1016/j.cub.2015.06.009.
Skerry, A.E., Saxe, R., 2014. A common neural code for perceived and inferred emotion. J. Neurosci. 34, 15997–16008. http://dx.doi.org/10.1523/JNEUROSCI.1676-14.2014.
Tamir, D.I., Thornton, M.A., Contreras, J.M., Mitchell, J.P., 2016. Neural evidence that three dimensions organize mental state representation: rationality, social impact, and valence. Proc. Natl. Acad. Sci. 113, 194–199. http://dx.doi.org/10.1073/pnas.1511905112.
Tanaka, K., 1993. Neuronal mechanisms of object recognition. Science 262, 685–688. http://dx.doi.org/10.1126/science.8235589.
Tervo, D.G.R., Tenenbaum, J.B., Gershman, S.J., 2016. Toward the neural implementation of structure learning. Curr. Opin. Neurobiol. 37, 99–105. http://dx.doi.org/10.1016/j.conb.2016.01.014.
Ünal, E., Papafragou, A., 2016. Production–comprehension asymmetries and the acquisition of evidential morphology. J. Mem. Lang. 89, 179–199. http://dx.doi.org/10.1016/j.jml.2015.12.001.
Winecoff, A., Clithero, J.A., Carter, R.M., Bergman, S.R., Wang, L., Huettel, S.A., 2013. Ventromedial prefrontal cortex encodes emotional value. J. Neurosci. 33, 11032–11039. http://dx.doi.org/10.1523/JNEUROSCI.4317-12.2013.
Young, L., Nichols, S., Saxe, R., 2010. Investigating the neural and cognitive basis of moral luck: it's not what you do but what you know. Rev. Phil. Psych. 1, 333–349. http://dx.doi.org/10.1007/s13164-010-0027-y.