Research Article

Psychological Science
2017, Vol. 28(3) 307–319

DOI: 10.1177/0956797616682029
www.psychologicalscience.org/PS

Same Story, Different Story: The Neural Representation of Interpretive Frameworks

Yaara Yeshurun[1,2], Stephen Swanson[3], Erez Simony[1,2], Janice Chen[1,2], Christina Lazaridi[4], Christopher J. Honey[5], and Uri Hasson[1,2]

[1]Department of Psychology, Princeton University; [2]Princeton Neuroscience Institute, Princeton University; [3]Department of Computer Science, Princeton University; [4]Creative Writing Program, Lewis Center for the Arts, Princeton University; and [5]Department of Psychological & Brain Sciences, Johns Hopkins University

Abstract

Differences in people's beliefs can substantially impact their interpretation of a series of events. In this functional MRI study, we manipulated subjects' beliefs, leading two groups of subjects to interpret the same narrative in different ways. We found that responses in higher-order brain areas—including the default-mode network, language areas, and subsets of the mirror neuron system—tended to be similar among people who shared the same interpretation, but different from those of people with an opposing interpretation. Furthermore, the difference in neural responses between the two groups at each moment was correlated with the magnitude of the difference in the interpretation of the narrative. This study demonstrates that brain responses to the same event tend to cluster together among people who share the same views.

Keywords

narrative, interpretation, context, theory of mind, neuroimaging

Received 1/6/16; Revision accepted 11/10/16

People see the world through a lens of attitudes and beliefs (Anderson, Reynolds, Schallert, & Goetz, 1977; Bransford & Johnson, 1972). For example, when Dartmouth and Princeton football fans watched the same game, each group of fans thought the other team was playing in an aggressive and unfair way, to such an extent that it seemed like they saw two very different games (Hastorf & Cantril, 1954). In the present study, we tested how neural responses to a narrative were altered when we gave two groups of subjects opposing views and beliefs about the situation depicted in the narrative.

Understanding interactions between characters in a story activates many brain regions, including regions implicated in thinking about the mental states of other people (Adolphs, 2009; Fletcher et al., 1995; Mar, 2011). The mentalizing network—which overlaps with the default-mode network (DMN; Mars et al., 2012) and includes medial prefrontal cortex, precuneus, and bilateral angular gyrus—is a set of regions activated when people infer the mental states of others (Schurz, Radua, Aichhorn, Richlan, & Perner, 2014; Van Overwalle & Baetens, 2009). To study this neural system, most prior studies have contrasted mentalizing-related conditions with conditions in which mentalizing is difficult or impossible; examples of the latter include stories without the necessary context (Ames, Honey, Chow, Todorov, & Hasson, 2015; Maguire, Frith, & Morris, 1999; Martin-Loeches, Casado, Hernandez-Tamames, & Alvarez-Linera, 2008; St George, Kutas, Martinez, & Sereno, 1999), temporally scrambled stories (Lerner, Honey, Silbert, & Hasson, 2011; Yarkoni, Speer, & Zacks, 2008), physical descriptions (Dodell-Feder, Koster-Hale, Bedny, & Saxe, 2011; Mason & Just, 2011; Saxe & Powell, 2006), and instructions asking subjects to attend to *how* other people do something rather than *why* they do it (Spunt & Adolphs, 2014; Spunt & Lieberman, 2012b).

Corresponding Author:
Yaara Yeshurun, Princeton University, Princeton Neuroscience Institute, Washington Rd., Princeton, NJ 08544
E-mail: yaara@princeton.edu

Although the DMN is known to be involved in the processing of other people's mental states, it remains unclear whether patterns of activity within the mentalizing network differ when there are two equally coherent and plausible interpretations of the mental states of characters in a narrative. To investigate this, we used functional MRI to record neural activity from subjects who listened to an audio rendition of a 12-min short story written by J. D. Salinger. The story was designed by Salinger to be highly ambiguous, with at least two possible—and very different—interpretations, and each of the subjects in our study was strongly primed to adopt one or the other. Behavioral tests confirmed that each group of listeners interpreted the story in different ways, depending on the prior context they received. Neurally, we found that similarity of the neural responses within the DMN, as well as parts of the mirror neuron system (Van Overwalle & Baetens, 2009) and high-level language-comprehension regions (Hagoort, Hald, Bastiaansen, & Petersson, 2004; Tesink et al., 2009), was higher among subjects who interpreted the story in the same way than between those subjects and subjects who interpreted the story differently. Furthermore, we found that the magnitude of neural distance across interpretations in the majority of these regions was significantly correlated with the magnitude of the difference in how the story was interpreted. This study demonstrates that shared understanding elicits shared neural response, within and outside of the DMN.

Method

Subjects

Forty-six right-handed subjects participated in the study. Six subjects were discarded from the analysis: 4 because of excessive head motion (> 2 mm) and 2 because they failed a stimulus-comprehension test (< 70% correct on a two-alternative forced-choice test). Twenty subjects (age: M = 20.85 years, SD = 3.73; 10 females, 10 males) were assigned to the cheating condition, and 20 other subjects (age: M = 21.45 years, SD = 3.42; 9 females, 11 males) were assigned to the paranoia condition. This sample size was chosen on the basis of previous studies in our lab that tested for similarities and differences in neural responses to naturalistic stimuli (Ames et al., 2015; Lerner et al., 2011). Experimental procedures were approved by the Princeton Institutional Review Board for Human Subjects. All subjects provided written informed consent.

Stimuli and experimental design

Subjects listened to an adapted version of the J. D. Salinger short story "Pretty Mouth and Green My Eyes." The adapted version was shorter than the original and included some sentences that were not present in the original text. It was read by a professional actor and ran 11 min, 32 s. The story was preceded by 18 s of neutral music and 3 s of silence, and it was followed by an additional 15 s of silence. These music and silence periods were discarded from all analyses. The story is about a phone conversation between two friends, Arthur and Lee. Arthur has returned home after a party after losing track of his wife, Joanie. He is calling Lee to share his concerns over her whereabouts. Lee is at home, and a woman is lying on the bed next to him. The woman's identity is ambiguous—she may or may not be Joanie, Arthur's wife. To disambiguate the story, we provided subjects with two different brief introductions (contexts) pointing toward two different interpretations. In the cheating condition, the context specified that Arthur's wife is cheating on him with Lee (sentences that differed between the two conditions are printed in italics):

> It is late at night and the phone is ringing. On one end of the line is Arthur; Arthur just came home from a party. He left the party without finding his wife, Joanie. *As always, Joanie was flirting with everybody at the party. Arthur is very upset.* On the other end is Lee, Arthur's friend. *He is at home with Joanie, Arthur's wife. Lee and Joanie have just returned from the same party. They have been having an affair for over a year now. They are thinking about the excuse Lee will use to calm Arthur this time.*

The paranoia context specified that Arthur is paranoid and that his wife is not cheating on him:

> It is late at night and the phone is ringing. On one end of the line is Arthur; Arthur just came home from a party. He left the party without finding his wife, Joanie. *As always, Arthur is paranoid, worrying that she might be having an affair, which is not true.* On the other end is Lee, Arthur's friend. *He is at home with his girlfriend, Rose. Lee and Rose have just returned from the same party, and are desperate to go to sleep. They do not know anything about Joanie's whereabouts, and are tired of dealing with Arthur's overreactions.*

The two contexts were intended to affect the interpretation of the characters' beliefs and emotions throughout the story. Story comprehension and the effect of context on story interpretation were assessed using a questionnaire at the conclusion of the experiment (outside the scanner; Fig. 1a).

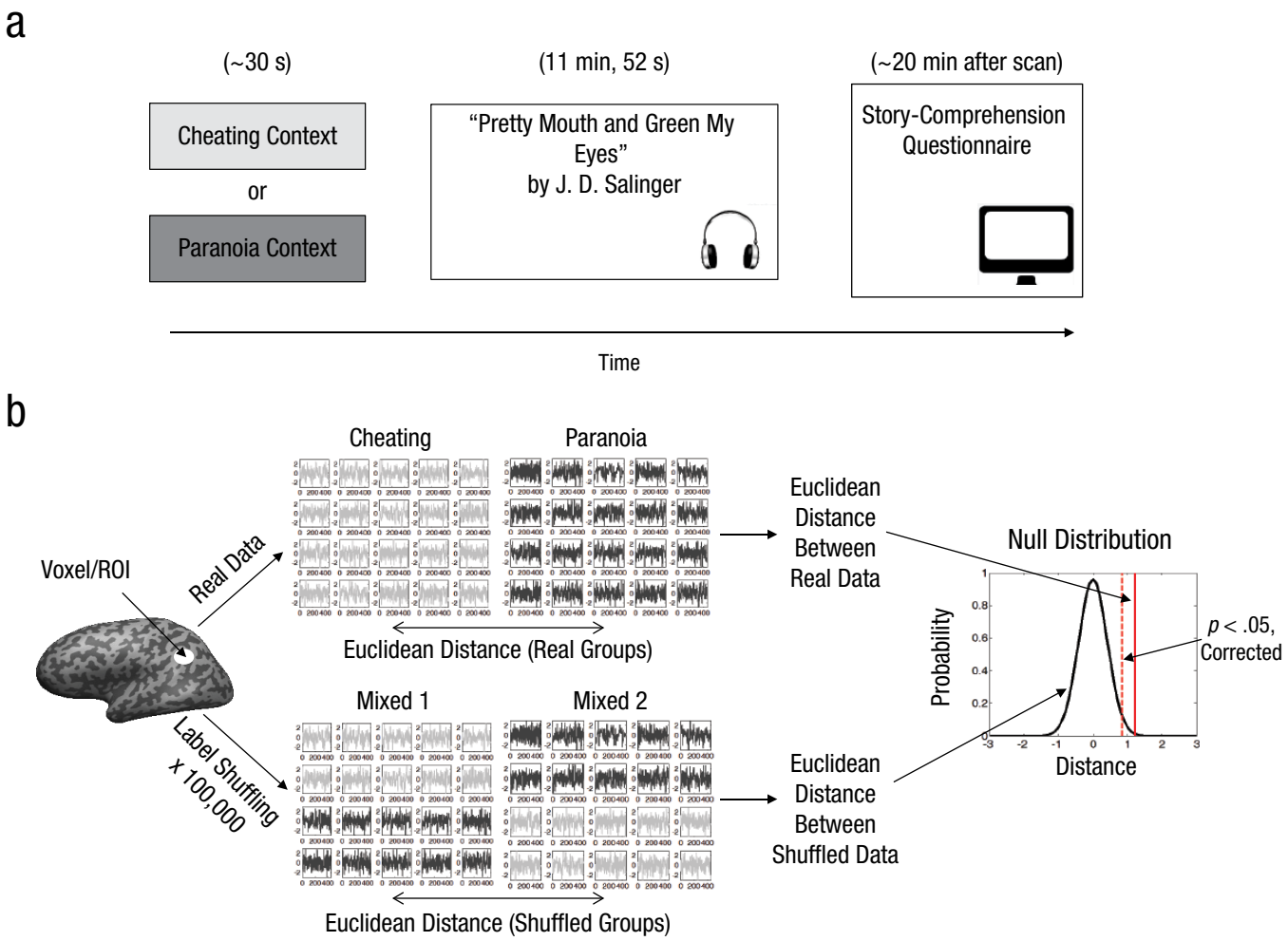

Fig. 1. Experimental procedure and design. Prior to listening to a recording of a short story in the scanner (a), subjects were placed in two groups: In one, subjects were primed to interpret the story as being about a wife cheating on her husband; in the other, subjects were primed to interpret the story as being about a husband being paranoid that his faithful wife is actually cheating on him. After listening to the story, subjects completed a story-comprehension questionnaire outside the scanner. In each voxel, we calculated the mean response of the 20 subjects in the cheating condition and the mean response of the 20 subjects in the paranoia condition (b), and then calculated the Euclidean distance between the activations in these groups' time courses. To test whether this distance was significant, we bootstrapped a null distribution (100,000 times) by calculating the Euclidean distance between two randomly sampled pseudogroups. ROI = region of interest.

Behavioral assessment

Immediately following the scan, each subject's comprehension of the story was assessed using a computerized questionnaire. Thirty-nine questions were presented, followed by a free-recall period and two forced-choice questions. Twenty-seven questions were context independent (e.g., "How many years were Arthur and Joanie together?"), and 12 questions were context dependent (e.g., "Why do you think Lee reacted that way?"). Two-tailed Student's t tests ($\alpha = .05$) on the forced-choice questions were conducted between the two conditions to evaluate the difference in subjects' comprehension.

MRI acquisition

Subjects were scanned in a 3T full-body MRI scanner (Skyra, Siemens, Erlangen, Germany) with a 12-channel head coil. For functional scans, images were acquired using a T2*-weighted echo-planar imaging pulse sequence—repetition time (TR) = 1,500 ms, echo time (TE) = 28 ms, flip angle = 64°—with each volume comprising 27 slices of 4-mm thickness with a 0-mm gap; slice-acquisition order was interleaved. In-plane resolution was 3×3 mm^2—field of view (FOV) = 192×192 mm^2. Anatomical images were acquired using a T1-weighted magnetization-prepared rapid-acquisition gradient echo

(MPRAGE) pulse sequence—TR = 2,300 ms, TE = 3.08 ms, flip angle = 9°, resolution = 0.89 mm^3, FOV = 256 mm^2. To minimize head movement, we stabilized subjects' heads with foam padding. Stimuli were presented using the Psychophysics Toolbox (Version 3.0.10; Pelli, 1997). Subjects were provided with MRI-compatible in-ear mono earbuds (Model S14, Sensimetrics, Malden, MA), which provided the same audio input to each ear. MRI-safe passive noise-canceling headphones were placed over the earbuds for noise removal and safety.

Data analysis

Our experiment was a context manipulation that altered the interpretation of the story. The context-dependent change in interpretation was not uniform across the whole story. To quantify this difference in a relatively objective way, we asked five raters (three females, two males) to independently analyze the interaction between the context and Salinger's text. The text was divided into 179 segments (mean duration = 3.77 s, *SD* = 2.39 s) by an expert independent annotator. The raters were instructed to rate how differently they thought that subjects in the cheating condition and in the paranoia condition would interpret each of these segments. Ratings were made on a scale from 1 to 5, and raters assessed three different aspects of interpretation: (a) beliefs of the characters, (b) emotions of the characters, and (c) intentions of the characters. Cronbach's αs for the ratings were .77 (beliefs), .74 (emotions), and .85 (intentions). We *z*-scored the ratings within a rater, within an aspect. Next, we obtained the mean rating of all raters in each segment. In segments in which the standard deviation was larger than 1 (belief: 16% of the segments; emotion: 18%; intention: 12%), we took the smallest normalized rating for the specific segment because in cases with substantial disagreement, we wanted to adopt a conservative stance. We preferred to underestimate rather than overestimate our context manipulation. The distribution of the average ratings across time was as follows: beliefs (2.88, *SD* = 1.09), emotions (3.13, *SD* = 0.95), and intentions (2.74, *SD* = 1.23).

Imaging analysis

Preprocessing. Functional MRI (fMRI) data were reconstructed and analyzed with the BrainVoyager QX software package (Brain Innovation, Maastricht, The Netherlands) and in-house software written in MATLAB (The MathWorks, Natick, MA). Preprocessing of functional scans included intrasession 3-D motion correction, slice-time correction, linear-trend removal, and high-pass filtering (two cycles per condition). Spatial smoothing was applied using a Gaussian filter of 6-mm full-width at half-maximum. The complete functional data set was transformed to 3-D Talairach space (Talairach & Tournoux, 1988). Hemodynamic delay was corrected on the basis of a correlation between the stimulus audio envelope and the blood-oxygen-level-dependent (BOLD) signal in early auditory areas (A1+) at the single-subject level. To align the neural data to the beginning of the story (and account for hemodynamic delay), we calculated for each subject the cross-correlation of the BOLD response in A1+ with the story audio envelope. We than shifted the BOLD response by the peak value of this cross-correlation (mean shift = 4.3 s, *SD* = 0.5 s).

Euclidean distance measure. We were interested in context-dependent differences in the neuronal responses of subjects listening to the same story. To test for such differences, we used a Euclidean distance metric voxel by voxel across the whole gray matter. For each voxel, we calculated the mean response of the 20 subjects in the cheating condition and the mean response of the 20 subjects in the paranoia condition. This resulted in two time courses, one for each condition, each with 450 time points. Next, we calculated the Euclidean distance between the two time courses using the following equation:

$$D = \sqrt{\Sigma_t \left(c(t) - p(t)\right)^2}, \quad (1)$$

where $c(t)$ is the mean BOLD time course measured in the cheating condition, and $p(t)$ is the mean BOLD time course measured in the paranoia condition. This procedure was used to obtain a distance value for each voxel in the gray matter.

Testing the statistical significance of time-course distance. To test whether the distance value was significantly larger than would be expected by chance (Fig. 1b), we simulated a null distribution using a permutation method. The data (20 time courses each from the cheating and paranoia conditions) from a specific voxel were extracted, and the labels of the groups were shuffled randomly to create two new pseudogroups, with corresponding mean time courses $\tilde{c}(t)$ and $\tilde{p}(t)$. We calculated the Euclidean distance between the mean responses in the pseudogroups as follows:

$$D = \sqrt{\Sigma_t \left(\tilde{c}(t) - \tilde{p}(t)\right)^2}. \quad (2)$$

The procedure of label shuffling and computing a surrogate Euclidean distance value was repeated 100,000 times. Thus, we obtained a null distribution of 100,000 distance values for the null hypothesis that there was no difference in the response time course across members of

the two groups. The p values of the empirical distances were derived using the following formula: (number of null values larger than the real value + 1)/100,000. We corrected for multiple comparisons by controlling the false-discovery rate (FDR; Benjamini & Hochberg, 1995) of the distance map with a q criterion of .05.

Euclidean-distance-based classification. The previous analysis aimed to reveal voxels that showed differences at the group level. To test whether these voxels also reliably identified subjects as having been exposed to one or the other context, we trained a classifier using the Euclidean distance measure in each of the voxels that demonstrated an effect in the previous analysis. The classifier received a training set and a testing set. The training set contained 19 Context 1 time courses (450 time points) and 19 Context 2 time courses (450 time points). The testing set contained the remaining two time courses (one from Context 1 and one from Context 2). During training, the classifier calculated the mean (centroid) of the 19 Context 1 time courses (Mean 1) and the mean (centroid) of the 19 Context 2 time courses (Mean 2). The classifier labeled a testing set as "Context 1" if the Euclidean distance between that set and Mean 1 was smaller than the distance between that set and Mean 2. The classifier labeled a testing set as "Context 2" if the Euclidean distance between that set and Mean 2 was smaller than the distance between that set and Mean 1. We used a leave-two-out algorithm, executed 400 times (each time, 2 different subjects were left out) in each voxel. In each time, the classifier could be correct (1) or incorrect (0). The classifier accuracy in a specific voxel was the average of the 400 trials (a number between 0 and 1).

Testing the statistical significance of the classifier accuracy value. To test whether the classifier accuracy value was significantly larger than would be expected by chance, we simulated a null distribution using a permutation method. The data (20 time courses each from the cheating and paranoia conditions) from a specific voxel were extracted, and the labels of the groups were shuffled randomly to create two new pseudogroups. We then classified each subject to Context 1 or Context 2 using the same classification procedure that was applied to the empirical data. This procedure of label shuffling and classifying was repeated 10,000 times. Thus, we obtained a null distribution of 10,000 classifier accuracies under the null hypothesis. This distribution reflects the probability of the classifier achieving a classification rate by chance. The p values of the empirical distances were computed using the following formula: (number of null values larger than the real value + 1)/10,000. We corrected for multiple comparisons by controlling the FDR (Benjamini & Hochberg, 1995) of the distance map with a q criterion of .05.

Correlation between neural distance and text distance. Our experiment included a context manipulation that altered the interpretation of the story, mainly by changing subjects' interpretation of the characters' mental states. The context-dependent change in the mental states of the characters was not uniform across the whole story. We wanted to test whether there was a correlation between the difference in neural response and the difference in interpretations of the mental states of the characters. To that end, we calculated the neural distance across groups within each TR in the voxels that had different responses across the whole story. We did this by taking the absolute value of the difference in the averaged BOLD response between the cheating and paranoia conditions at each time point. This resulted in a 2,009 (voxels) × 450 (TRs) matrix of difference in neural response values. Next, within each voxel, we correlated distance in neural response (D; see Equation 1) with difference in the beliefs, emotions, and intentions attributed to the characters. Statistical significance of each correlation coefficient (whether each was significantly different from zero) was computed using a bootstrapping procedure. For every empirical-distance time course in every voxel, 10,000 bootstrapped time series were generated using a phase-randomization procedure, which preserves the temporal autocorrelation in the distance time series. Phase randomization was performed by fast-Fourier-transforming the signal, randomizing the phase of each Fourier component, and then inverting the Fourier transformation. This procedure leaves the power spectrum of the signal unchanged but removes temporal alignment of the signals. Using these bootstrapped time courses, a null distribution of the correlation values was determined for each of the 2,009 voxels. The p values of the empirical correlations were computed by comparison with these null distributions. We corrected for multiple comparisons by controlling the FDR (Benjamini & Hochberg, 1995) of the distance map with a q criterion of .05.

Euclidean distance measure in the mentalizing network. We ran all analyses at the whole-brain, voxel-by-voxel level as well as on a predefined set of regions of interest (ROIs). We defined ROIs using three classical localizers for the mentalizing system: the false-belief localizer, the why-versus-how theory-of-mind localizer, and the Neurosynth mini-meta-analysis localizer.

False-belief localizer. After participating in our main experiment (listening to "Pretty Mouth and Green My Eyes"), subjects were presented with the belief-versus-nonbelief localizer (Dodell-Feder et al., 2011). We created a 40-subject general linear model and defined nine ROIs on the basis of the belief > photo contrast used by Dodell-Feder and colleagues, q(FDR) < .01. We selected a

cube of 10 × 10 × 10 voxels around the peak coordinates. The ROIs were right and left temporoparietal junction (TPJ), precuneus, posterior cingulate cortex, ventromedial prefrontal cortex (vmPFC), dorsomedial prefrontal cortex (dmPFC), right middle frontal gyrus (MFG), right superior temporal sulcus (STS), and right temporal pole (see Table S1 in the Supplemental Material).

Why-versus-how localizer. We downloaded the *t*-statistic map (http://neurovault.org/images/3078/) for the why > how contrast in the why-versus-how localizer developed by Spunt and colleagues (Spunt & Adolphs, 2014; Spunt & Lieberman, 2012b). We took a cube of 10 × 10 × 10 voxels around the peak coordinates (after converting them from Montreal Neurological Institute, MNI, to Talairach coordinates), also using our gray-matter mask derived from the 40 subjects. Our 10 ROIs were right and left TPJ, posterior cingulate cortex, vmPFC, dmPFC, right and left inferior frontal gyrus (IFG), left superior temporal gyrus (STG), right temporal pole, and right hippocampus (see Table S1).

Neurosynth theory-of-mind localizer. We generated a mini meta-analysis of theory-of-mind regions using Neurosynth, a platform developed by Tal Yarkoni, for large-scale, automated synthesis of fMRI data from multiple studies. The mini meta-analysis contained 140 studies discovered by searching for the words "theory mind" (http://www.neurosynth.org/analyses/terms/theory%20mind/). We selected a cube of 10 × 10 × 10 voxels around the peak coordinates (after converting them from MNI to Talairach coordinates), also using our 40-subjects gray-matter mask. Our 12 ROIs were right and left TPJ, precuneus, posterior cingulate cortex, vmPFC, dmPFC, right and left IFG, right and left STS, and right and left temporal pole (see Table S1). We tested whether the neural response in these ROIs differed significantly between the cheating and the paranoia conditions using the Euclidean distance measure.

Results

Behavioral results: similar comprehension, different interpretation

To assess general comprehension of the story, we first tested subjects on context-independent questions (e.g., "What was the girl doing when the phone rang?"; possible answers: "Lying on the bed" or "Sitting in a chair"). The correct answer to these questions was not dependent on the context presented to the subjects. Indeed, comprehension of the story was high (cheating: M = 93.88% correct, SD = 1.4%; paranoia: M = 93.52% correct, SD = 1.0%), with no significant difference between the two conditions, $t(38) = 0.22$, $p = .83$ (Fig. 2). We then tested whether the context manipulation changed subjects' interpretation of the story using context-dependent questions (e.g., "Why do you think Lee didn't want Arthur to come over?"). The correct answers to these questions were dependent on the context. Subjects in the cheating condition interpreted the characters' intentions differently and chose a different answer ("He did not want him to find out that his wife is there") from subjects in the paranoia condition ("He was desperate to go to sleep"). Indeed, we found a significant difference between conditions for these interpretation questions, $t(38) > 5.1$, $p < 10^{-10}$ (Fig. 2). Subjects in the cheating condition had an average of 80.83% (SD = 3.1%) cheating-appropriate answers, while subjects in the paranoia condition had an average of 73.75% (SD = 6.77%) paranoia-appropriate answers.

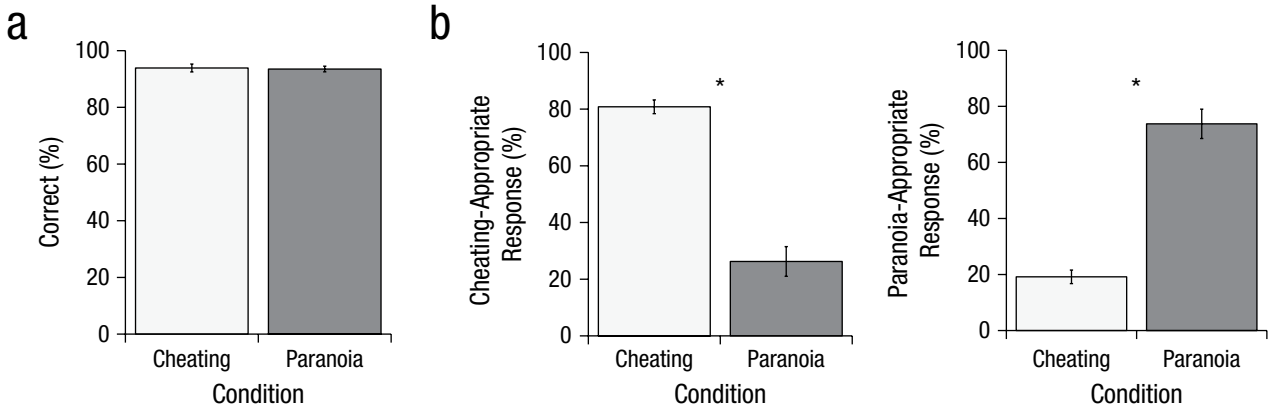

Fig. 2. Behavioral results demonstrating the context effect. The graph in (a) shows the mean percentage of correct responses to context-independent questions (i.e., that did not depend on subjects' assignment to condition). The graphs in (b) show the mean percentage of responses in each condition that were appropriate to the cheating context (left) and paranoia context (right). Asterisks indicate significant differences between conditions ($p < .05$). Error bars represent ±1 *SE*.

Thus, the same story was interpreted in coherent yet different ways depending on the preceding context. This design is different from that used in most context-based studies, which focus on the dissociation between situations in which changes in context creates changes in the coherence of the text (e.g., Ames et al., 2015; Lerner et al., 2011; Maguire et al., 1999; Yarkoni et al., 2008).

Context-dependent responses in extensive brain areas

We were interested in identifying brain regions in which the neural response was modulated by the interpretation of the narrative. To that end, we calculated the Euclidean distance between the time courses of the two conditions in each voxel across the whole gray matter (see Fig. 1b). Figure 3a shows the mean responses and the Euclidean distance between the responses for one voxel in precuneus and one in right primary auditory cortex. We observed no systematic differences in the responses in the auditory cortex across the two conditions; in contrast, the neural distance of the mean activity between the two conditions was significantly different in the precuneus. Performing this analysis voxel by voxel, we observed a significant difference in the neural responses between the two conditions in many brain regions (2,009 voxels), including most of the mentalizing network (right and left TPJ, precuneus, right MFG, vmPFC), the right and left hippocampus, language-related areas in the ventrolateral prefrontal cortex (vlPFC) and anterior IFG, and areas related to the mirror neuron system in the right and left premotor cortex (PMC; Fig. 3b and Table 1). Systematic difference in the neural responses in the two conditions was also observed when we performed the analysis using three sets of ROIs defined by three independent theory-of-mind localizers (Fig. 3c; also see Table S1).

How many of these voxels reliably identified the context at the single-subject level? We trained a classifier using the Euclidean distance measure in each of the 2,009 voxels that demonstrated an effect in our group analysis. We discovered that out of all voxels that showed significantly different activation between the cheating and the paranoia conditions at the group level, 72% (1,446) could be used for successful classification at the single-subject level. The binary classification performance ranged from 66% correct in some voxels to 88% correct in other regions, with the highest mean classification performance observed for voxels in the right TPJ, bilateral precuneus, and bilateral vlPFC (Fig. 4). Thus, the response time courses in these regions were more similar in subjects who held the same interpretation than in subjects who held different interpretations.

Correlation of the neural responses in differentiating voxels with changes in the interpretation of characters' beliefs, emotions, and intentions

Which aspects of the interpretation (if any) drive the context-dependent changes in response time courses within these differentiating voxels? To answer this question, we correlated differences in the interpretation of the characters' beliefs, emotions, and intentions (Fig. 5a) with distances in neural response (Fig. 5b; we tested whether the correlation coefficients of the aspects were significantly different from zero, not whether they were significantly different from each other; see Method for details). We found that (a) the distance in the neural response of 785 of the 2,009 differentiating voxels was correlated with differences in the beliefs of the characters (minimum r = .13, maximum r = .34; M = .23, SD = .04), (b) the distance in neural response of 968 of the 2,009 differentiating voxels was correlated with differences in the emotions of the characters (minimum r = .11, maximum r = .34; M = .22, SD = .04), and (c) the distance in neural response in none of the 2,009 differentiating voxels was correlated with differences in the intentions of the characters.

When we mapped these voxels on the brain, we found that differences in the response of voxels within the precuneus, right TPJ, and dmPFC were correlated with difference in the beliefs and emotions attributed to the characters (Fig. 5c). Other regions, outside of the DMN, also correlated with differences in the interpretation of different aspects. Specifically, changes in neural representation within right PMC, within left and right vlPFC, and within anterior left STS were significantly correlated with interpretations of the characters' beliefs and emotions (Fig. 5c), and changes in the neural representation within left and right hippocampus were significantly correlated with interpretation of the characters' emotions (Fig. 5c).

It is important to note that there were regions where activations discriminated between the cheating and paranoia conditions, but these regions were not correlated with changes in any of the aspects measured. These regions included medial left STS, left PMC, inferior left TPJ, and parts of the inferior and dorsal precuneus (Fig. 5c). This suggests that these regions reflected context-dependent differences that are not related to the interpretation of characters' intentions, emotions, or beliefs.

Discussion

The way people interpret events is dependent both on the actual external input and on internal cognitive information (Anderson et al., 1977; Bransford & Johnson, 1972).

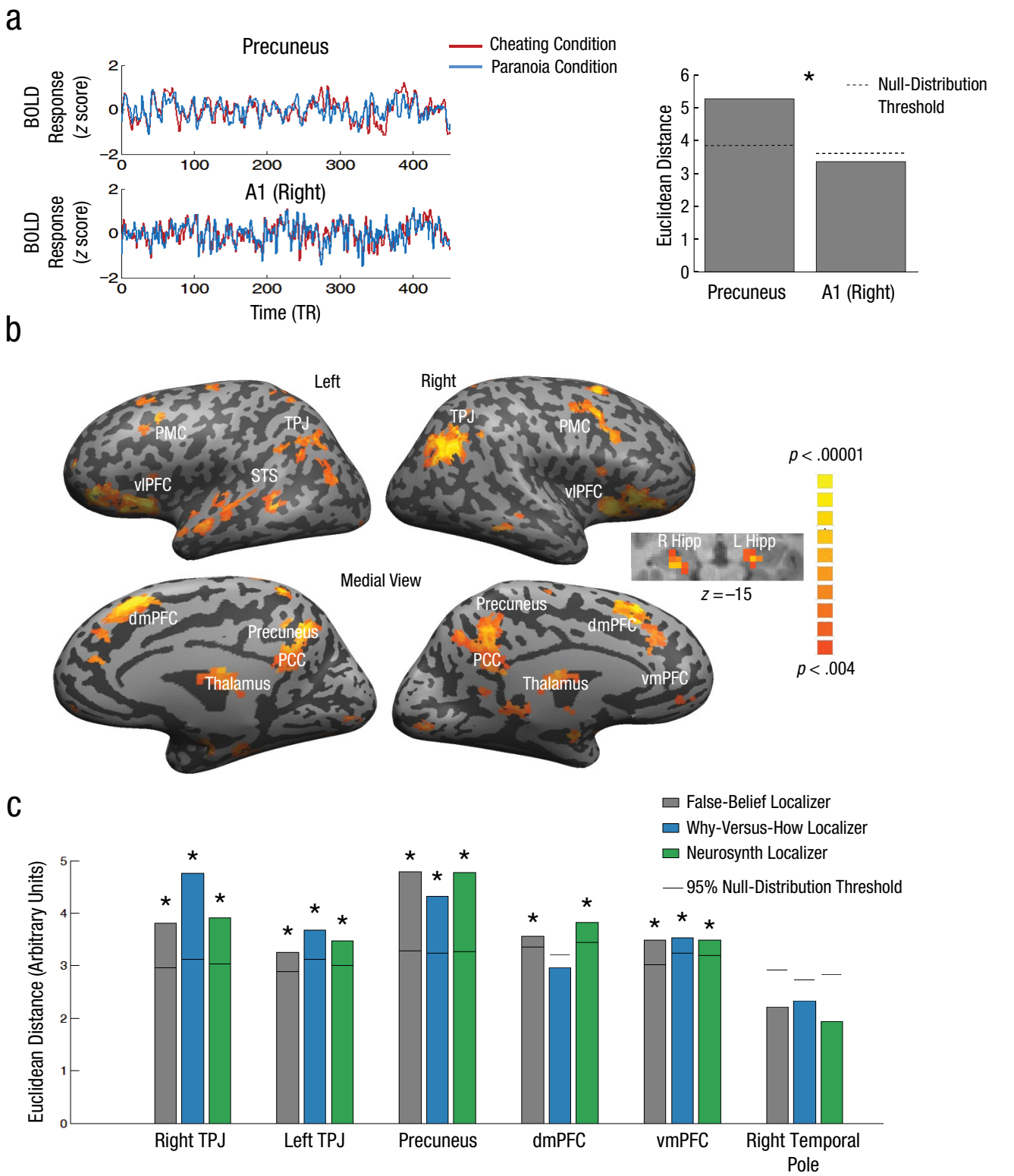

Fig. 3. Neural results demonstrating the context effect. Example mean blood-oxygen-level-dependent (BOLD) repetition-time (TR) courses for subjects in the cheating and paranoia conditions (a) were sampled from one voxel in the precuneus and one voxel in the right auditory cortex (A1). The bar graph shows the Euclidean distance between the two time courses, separately for each region. The asterisk indicates a significant difference between the two conditions ($p < .05$), as indicated by the null distribution. The Euclidean distance maps across the whole brain (b) show regions in which activations differed significantly between conditions (minimum cluster size > 10 mm^2; p values are false-discovery-rate corrected). Euclidean distance between the two conditions within mentalizing-network regions of interest (c) were defined by the false-belief localizer, the why-versus-how localizer, and the Neurosynth localizer. For each localizer, asterisks indicate significant differences between conditions ($p < .05$, false-discovery-rate corrected), as indicated by the null-distribution threshold. dmPFC = dorsomedial prefrontal cortex, Hipp = hippocampus, PCC = posterior cingulate cortex, PMC = premotor cortex, STS = superior temporal sulcus, TPJ = temporoparietal junction, vlPFC = ventrolateral prefrontal cortex, vmPFC = ventromedial prefrontal cortex.

Table 1. Brain Regions With Significant Neural Distance Between Activations in the Cheating and the Paranoia Conditions

Region	Hemisphere	Peak *t*	Coordinates			Number of voxels
			x	*y*	*z*	
Precuneus	Bilateral	7.5795	5	−53	33	177
Posterior cingulate cortex	Bilateral	5.3025	−1	−50	18	50
Temporoparietal junction	Right	7.9815	44	−56	31	205
Temporoparietal junction	Left	4.5300	−41	−59	24	79
Superior temporal sulcus	Left	5.1105	−52	−7	−6	107
Temporal pole	Left	4.6965	−52	10	−12	25
Premotor cortex	Right	5.7240	38	11	40	139
Premotor cortex	Left	5.3805	−37	10	40	105
Dorsomedial prefrontal cortex	Bilateral	8.3085	3	19	50	115
Ventromedial prefrontal cortex	Bilateral	3.8610	−6	45	−3	33
Ventrolateral prefrontal cortex	Right	9.9750	43	31	−3	268
Ventrolateral prefrontal cortex	Left	8.6250	−43	42	−4	207
Hippocampus	Right	5.1930	20	−12	−19	22
Hippocampus	Left	5.1015	−19	−12	−15	28
Thalamus	Bilateral	6.9105	3	−20	12	41

Note: Coordinates are shown for the local maxima. Coordinates are given in Talairach space, where *x*, *y*, and *z* refer to the left-right, anterior-posterior, and inferior-superior dimensions, respectively.

By keeping the external input constant and changing the context in which subjects understood that input, we identified networks of brain regions in which neural responses clustered together as a function of interpretation. These regions included the mentalizing network, part of the mirror neuron system, and part of the story-comprehension network (Jung-Beeman, 2005). Further, we found a correlation between the magnitude of the difference in the neural response and the magnitude of the difference in externally assessed aspects of the interpretation.

Previous studies have identified the mentalizing network involved in thinking about characters' mental states (Mar, 2011; Saxe & Powell, 2006; Schurz et al., 2014; Spunt & Adolphs, 2014) by comparing mentalizing tasks and nonmentalizing tasks (i.e., thinking about physical vs. mental causations, or thinking about why vs. how a

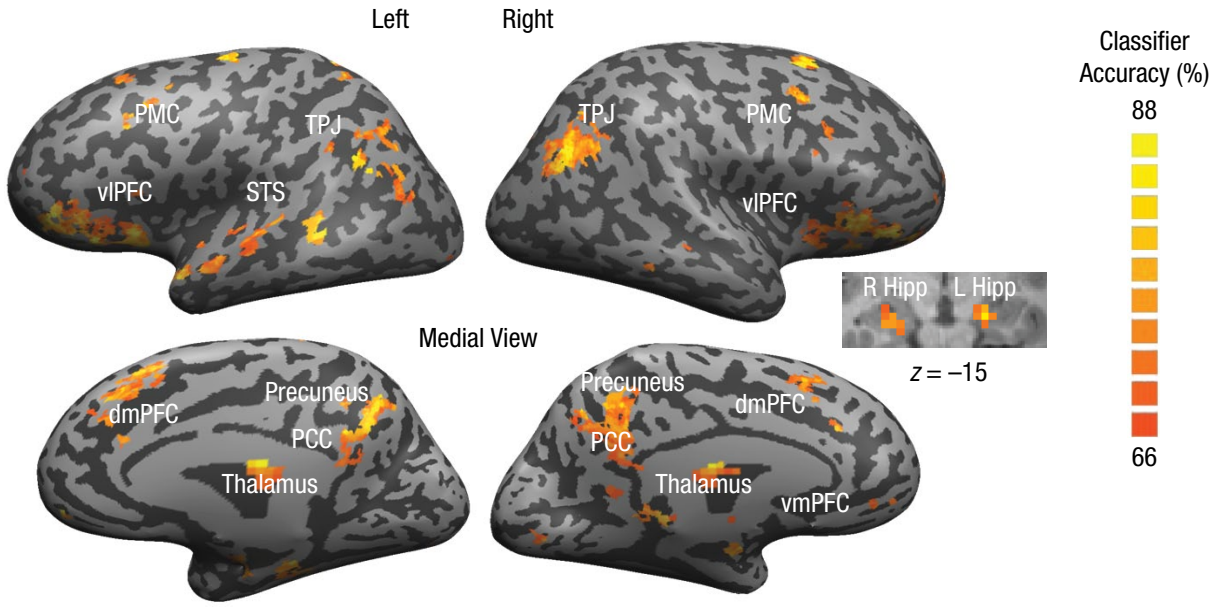

Fig. 4. Map of classification accuracy within the voxels in which there was a significant distance between activations in the two conditions. dmPFC = dorsomedial prefrontal cortex, Hipp = hippocampus, PCC = posterior cingulate cortex, PMC = premotor cortex, STS = superior temporal sulcus, TPJ = temporoparietal junction, vlPFC = ventrolateral prefrontal cortex, vmPFC = ventromedial prefrontal cortex.

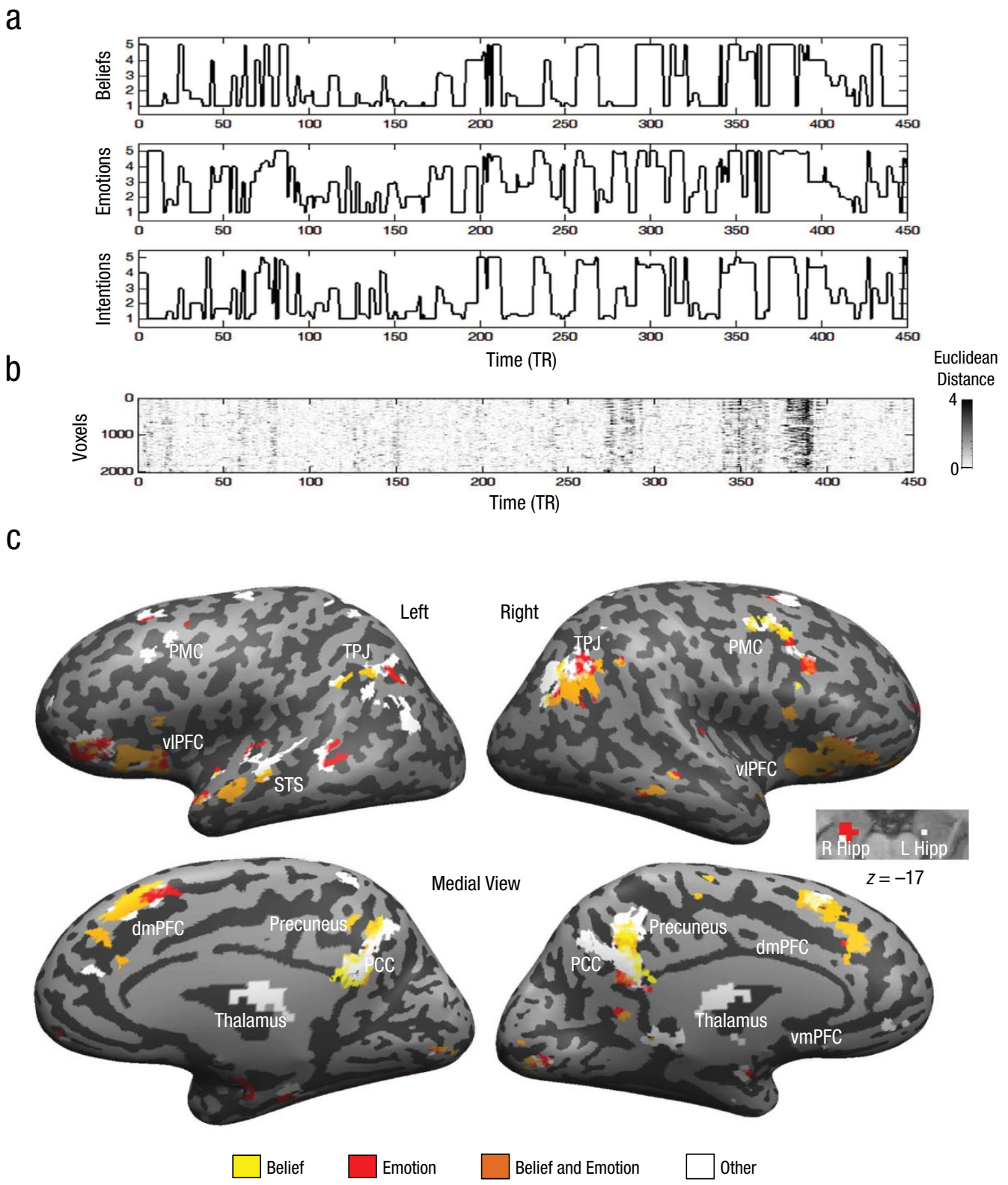

Fig. 5. Correlation between the difference in interpretation and the difference in neural response. Ratings of the difference between the cheating and paranoia conditions in the beliefs, emotions, and intentions attributed to the characters (a) are shown as a function of repetition time (TR; i.e., 1.5 s) during the audio recording. The Euclidean distance between voxels in which activation significantly differed between the two conditions (b) is shown as a function of TR. The brain maps (c) show voxels with correlations between neural differences and differences in the beliefs of the characters, emotions of the characters, and both. Colored regions were significant ($p < .05$, false-discovery-rate corrected); regions with a nonsignificant correlation are marked in white. dmPFC = dorsomedial prefrontal cortex, Hipp = hippocampus, PCC = posterior cingulate cortex, PMC = premotor cortex, STS = superior temporal sulcus, TPJ = temporoparietal junction, vlPFC = ventrolateral prefrontal cortex.

person is performing an action). Some other studies controlled the level of complexity needed for the interpretation of characters' intentions: For example, ironic statements can be more taxing, in terms of computational resources, than face-value sentences (Bašnáková, Weber, Petersson, van Berkum, & Hagoort, 2014; Uchiyama et al., 2012), and ambiguous inferences are more demanding then unambiguous inferences (Jenkins & Mitchell, 2010). These studies found increased activity in regions within the mentalizing network as a function of the mentalizing load. In many cases, however, the mental states attributed to another person vary not only in the level of complexity, but also in the attributed content.

To address this concern, recent studies have explored neural representations of different attributed content (Tamir, Thornton, Contreras, & Mitchell, 2016): for example, distinguishing between harmful and impure acts (Chakroff et al., 2016), between cooperative or competitive behavior of other people (Tsoi, Dungan, Waytz, & Young, 2016), between preferences and beliefs (Jenkins & Mitchell, 2010), and between different emotional states attributed to other people (Skerry & Saxe, 2015). These elegant studies demonstrated a means of discriminating between different mental states, but in all of them, the different mental states were generated by manipulating the actual content of the stimuli across conditions; this is in contrast to our study, in which the stimulus (a 12-min real-life narrative) was identical across the two groups. Holding the stimulus fixed while manipulating subjects' mental states is of vast importance, given that in many real-life situations, people can markedly differ in their assessment of the mental states of others. For example, depending on the context (e.g., your political views), the same statement (expressed in a political debate by one of the candidates) might be interpreted as an expression of sincere concern or as a condescending remark. We are aware of only one other study that classified the content of mental states (discriminating between an "intentional harm" and an "accidental harm") generated by the exact same situation, on the basis of brain patterns within the right TPJ (Koster-Hale, Saxe, Dungan, & Young, 2013). Our findings are congruent with this result and extend it in three different ways.

First, whereas in the aforementioned studies, subjects were explicitly instructed to consider and judge other people's intentions and beliefs, the use of a real-life narrative enabled us to examine these processes occurring spontaneously as listeners comprehend a narrative. Second, we searched for differences in neural representations across the whole brain. We found that neural activity not only in the right TPJ, but also in the majority of DMN regions, was sufficient for distinguishing the two interpretations of the narrative (Figs. 3 and 4). And finally, the story stimulus allowed us to dissociate mental states associated with different types of content. In particular, we found that the change in the response in some of these regions was correlated with changes in the interpretation of character's emotions and beliefs (Fig. 5c).

We found interpretation-based shared responses across subjects not only in the mentalizing network, but also in the premotor cortex (Fig. 2c), a part of the mirror neuron system (Rizzolatti, Fogassi, & Gallese, 2001). Notably, we also found that the difference in the response of the right premotor cortex was correlated with differences in the interpretation of the characters' beliefs and emotions (Fig. 5). The role of the mirror neuron system in thinking about the intentions and beliefs underlying a specific behavior has been extensively discussed (Gallese, Keysers, & Rizzolatti, 2004; Saxe, 2005; Van Overwalle & Baetens, 2009). Advocates of simulation theory suggest that the mirror neuron system is required in order to simulate the external signs of the mental state (e.g., smiling or reaching for a cup) and that this information is then used by the brain to understand the mental state underlying these external signs (happy or thirsty, respectively; Keysers & Gazzola, 2006). For example, the mirror neuron system exhibited increased activity when subjects viewed social facial expressions relative to nonsocial facial movements (Montgomery, Seeherman, & Haxby, 2009) and social animations of rigid geometric shapes (Gobbini, Koralek, Bryan, Montgomery, & Haxby, 2007). Further, it was sensitive to the intentions behind the same motor action (Iacoboni et al., 2005). In our experiment, the description of the behavior was identical (a telephone conversation between two friends), whereas the interpretation of the overt behavior was different (the friends' beliefs, emotions, and intentions). Our finding that part of the mirror neuron system differentiated between the cheating and paranoia contexts is the first evidence of its ability to differentiate between the intentions and beliefs underlying nonmotor behaviors. It also suggests a mixed model in which both mental simulation and motor simulation work in unison (Saxe, 2005; Spunt & Lieberman, 2012a) to interpret characters' beliefs and intentions in the story.

In addition to the DMN and the mirror neuron system, we found that the right and left vlPFC (anterior IFG and orbitofrontal cortex) were modulated by the context-driven interpretation (Fig. 3b). The more posterior part of this vlPFC activation—bilateral anterior IFG—plays an important role in language comprehension (Jung-Beeman, 2005). It has been shown to be involved in integration of world knowledge and local context with the presented text (Hagoort et al., 2004; Tesink et al., 2009) and to be sensitive to inconsistencies in emotional or chronological information (Ferstl, Rinck, & von Cramon, 2005). The current study extends these findings by showing that (a) vlPFC is involved in integration of previous knowledge with the text, even when two different contexts are both consistent with the text, and (b) this region activation is

modulated by interpretation. Our results are consistent with the proposal that vlPFC is part of a system that facilitates the construction of knowledge about people by relating past experiences with the personality of the person (Ranganath & Ritchey, 2012).

We found that shared understanding elicits shared neural responses in the mentalizing network and other higher-order regions: Responses of a listener who had one interpretation of the characters' actions (e.g., that the friend is lying in the cheating condition) were more similar to responses in other listeners who had a similar interpretation and different from the responses of listeners with a different interpretation (e.g., that the husband is paranoid and the friend is exhausted). This allowed us to predict the interpretation of one subject by correlating his or her neural responses with those of other subjects who have common prior knowledge (see Fig. 4). These results are in line with those of other studies demonstrating that instructions to focus on different aspects of the narrative, explicitly (Cooper, Hasson, & Small, 2011) or by changing the perspective taken (Lahnakoski et al., 2014), resulted in different patterns of response in several brain regions. However, in both prior studies, the instructions had a large effect on the low-level information extracted from the narrative (e.g., inducing different eye-movement patterns across groups). Our results demonstrate that a change in the prior context (four lines of text before the beginning of the story) did not change listeners' general comprehension of the narrative (Fig. 2) but resulted in interpretation-based group-selective neural alignment while processing that narrative. In real life, when different groups accumulate the same knowledge of the world from different sources (e.g., news channels with an opposing political orientation), there is likely even greater possibility for in-group clustering and out-group differentiation of neural activity. Thus, people can gradually construct distinct microworlds in their brains by integrating idiosyncratic prior knowledge and beliefs with shared information about events. Here, we have characterized the neural divergence of people who interpret identical input through the lens of different beliefs.

Action Editor

Ralph Adolphs served as action editor for this article.

Author Contributions

Y. Yeshurun and U. Hasson developed the study concept. Y. Yeshurun, C. J. Honey, and U. Hasson designed the study. Data were collected by Y. Yeshurun and J. Chen. Data were analyzed by Y. Yeshurun, S. Swanson, C. Lazaridi, and E. Simony. Y. Yeshurun and U. Hasson drafted the manuscript, and C. J. Honey and J. Chen provided critical revisions. All authors approved the final version of the manuscript for submission.

Acknowledgments

We would like to thank G. Dishon for his help in finding experimental materials and M. Nguyen and C. Baldassano for helpful comments on the manuscript.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This work was supported by National Institutes of Health Grant No. RO1-MH094480 (U. Hasson, Y. Yeshurun, E. Simony, and J. Chen), the Rothschild Foundation (Y. Yeshurun), and The Israel National Postdoctoral Program for Advancing Women in Science (Y. Yeshurun).

Supplemental Material

Additional supporting information can be found at http://journals.sagepub.com/doi/suppl/10.1177/0956797616682029

References

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

Ames, D. L., Honey, C. J., Chow, M. A., Todorov, A., & Hasson, U. (2015). Contextual alignment of cognitive and neural dynamics. *Journal of Cognitive Neuroscience*, *27*, 655–664.

Anderson, R. C., Reynolds, R. E., Schallert, D. L., & Goetz, E. T. (1977). Frameworks for comprehending discourse. *American Educational Research Journal*, *14*, 367–381.

Bašnáková, J., Weber, K., Petersson, K. M., van Berkum, J., & Hagoort, P. (2014). Beyond the language given: The neural correlates of inferring speaker meaning. *Cerebral Cortex*, *24*, 2572–2578.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society*, *57*, 289–300.

Bransford, J., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, *11*, 717–726.

Chakroff, A., Dungan, J., Koster-Hale, J., Brown, A., Saxe, R., & Young, L. (2016). When minds matter for moral judgment: Intent information is neurally encoded for harmful but not impure acts. *Social Cognitive and Affective Neuroscience*, *11*, 476–484.

Cooper, E. A., Hasson, U., & Small, S. L. (2011). Interpretation-mediated changes in neural activity during language comprehension. *NeuroImage*, *55*, 1314–1323.

Dodell-Feder, D., Koster-Hale, J., Bedny, M., & Saxe, R. (2011). fMRI item analysis in a theory of mind task. *NeuroImage*, *55*, 705–712.

Ferstl, E. C., Rinck, M., & von Cramon, D. Y. (2005). Emotional and temporal aspects of situation model processing during text comprehension: An event-related fMRI study. *Journal of Cognitive Neuroscience*, *17*, 724–739.

Fletcher, P. C., Happé, F., Frith, U., Baker, S. C., Dolan, R. J., Frackowiak, R. S., & Frith, C. D. (1995). Other minds in the

brain: A functional imaging study of "theory of mind" in story comprehension. *Cognition*, *57*, 109–128.

Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences*, *8*, 396–403.

Gobbini, M. I., Koralek, A. C., Bryan, R. E., Montgomery, K. J., & Haxby, J. V. (2007). Two takes on the social brain: A comparison of theory of mind tasks. *Journal of Cognitive Neuroscience*, *19*, 1803–1814.

Hagoort, P., Hald, L., Bastiaansen, M., & Petersson, K. M. (2004). Integration of word meaning and world knowledge in language comprehension. *Science*, *304*, 438–441.

Hastorf, A. H., & Cantril, H. (1954). They saw a game: A case study. *Journal of Abnormal and Social Psychology*, *49*, 129–134.

Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology*, *3*(3), Article e79. doi:10.1371/journal.pbio.0030079

Jenkins, A. C., & Mitchell, J. P. (2010). Mentalizing under uncertainty: Dissociated neural responses to ambiguous and unambiguous mental state inferences. *Cerebral Cortex*, *20*, 404–410.

Jung-Beeman, M. (2005). Bilateral brain processes for comprehending natural language. *Trends in Cognitive Sciences*, *9*, 512–518.

Keysers, C., & Gazzola, V. (2006). Towards a unifying neural theory of social cognition. *Progress in Brain Research*, *156*, 379–401.

Koster-Hale, J., Saxe, R., Dungan, J., & Young, L. L. (2013). Decoding moral judgments from neural representations of intentions. *Proceedings of the National Academy of Sciences, USA*, *110*, 5648–5653.

Lahnakoski, J. M., Glerean, E., Jääskeläinen, I. P., Hyönä, J., Hari, R., Sams, M., & Nummenmaa, L. (2014). Synchronous brain activity across individuals underlies shared psychological perspectives. *NeuroImage*, *100*, 316–324.

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, *31*, 2906–2915.

Maguire, E. A., Frith, C. D., & Morris, R. G. (1999). The functional neuroanatomy of comprehension and memory: The importance of prior knowledge. *Brain, 122*(Pt. 10), 1839–1850.

Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annual Reviews of Psychology*, *62*, 103–134.

Mars, R. B., Neubert, F.-X., Noonan, M. P., Sallet, J., Toni, I., & Rushworth, M. F. S. (2012). On the relationship between the "default mode network" and the "social brain." *Frontiers in Human Neuroscience, 6*, Article 189. doi:10.3389/fnhum.2012.00189

Martin-Loeches, M., Casado, P., Hernandez-Tamames, J. A., & Alvarez-Linera, J. (2008). Brain activation in discourse comprehension: A 3t fMRI study. *NeuroImage*, *41*, 614–622.

Mason, R. A., & Just, M. A. (2011). Differentiable cortical networks for inferences concerning people's intentions versus physical causality. *Human Brain Mapping*, *32*, 313–329.

Montgomery, K. J., Seeherman, K. R., & Haxby, J. V. (2009). The well-tempered social brain. *Psychological Science*, *20*, 1211–1213.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, *10*, 437–442.

Ranganath, C., & Ritchey, M. (2012). Two cortical systems for memory-guided behaviour. *Nature Reviews Neuroscience*, *13*, 713–726.

Rizzolatti, G., Fogassi, L., & Gallese, V. (2001). Neurophysiological mechanisms underlying the understanding and imitation of action. *Nature Reviews Neuroscience*, *2*, 661–670.

Saxe, R. (2005). Against simulation: The argument from error. *Trends in Cognitive Sciences*, *9*, 174–179.

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, *17*, 692–699.

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

Skerry, A. E., & Saxe, R. (2015). Neural representations of emotion are organized around abstract event features. *Current Biology*, *25*, 1945–1954.

Spunt, R. P., & Adolphs, R. (2014). Validating the why/how contrast for functional MRI studies of theory of mind. *NeuroImage*, *99*, 301–311.

Spunt, R. P., & Lieberman, M. D. (2012a). An integrative model of the neural systems supporting the comprehension of observed emotional behavior. *NeuroImage*, *59*, 3050–3059.

Spunt, R. P., & Lieberman, M. D. (2012b). Dissociating modality-specific and supramodal neural systems for action understanding. *Journal of Neuroscience*, *32*, 3575–3583.

St George, M., Kutas, M., Martinez, A., & Sereno, M. I. (1999). Semantic integration in reading: Engagement of the right hemisphere during discourse processing. *Brain, 122*(Pt. 7), 1317–1325.

Talairach, J., & Tournoux, P. (1988). *Co-planar stereotaxic atlas of the human brain* (M. Rayport, Trans.). New York, NY: Thieme Medical Publishers.

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. *Proceedings of the National Academy of Sciences, USA*, *113*, 194–199.

Tesink, C. M. J. Y., Petersson, K. M., van Berkum, J. J. A., van den Brink, D., Buitelaar, J. K., & Hagoort, P. (2009). Unification of speaker and meaning in language comprehension: An fMRI study. *Journal of Cognitive Neuroscience*, *21*, 2085–2099.

Tsoi, L., Dungan, J., Waytz, A., & Young, L. (2016). Distinct neural patterns of social cognition for cooperation versus competition. *NeuroImage*, *137*, 86–96.

Uchiyama, H. T., Saito, D. N., Tanabe, H. C., Harada, T., Seki, A., Ohno, K., . . . Sadato, N. (2012). Distinction between the literal and intended meanings of sentences: A functional magnetic resonance imaging study of metaphor and sarcasm. *Cortex*, *48*, 563–583.

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage*, *48*, 564–584.

Yarkoni, T., Speer, N. K., & Zacks, J. M. (2008). Neural substrates of narrative comprehension and memory. *NeuroImage*, *41*, 1408–1425.