

# Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence

Diana I. Tamir<sup>a,1</sup>, Mark A. Thornton<sup>b,1,2</sup>, Juan Manuel Contreras<sup>c</sup>, and Jason P. Mitchell<sup>b</sup>

<sup>a</sup>Department of Psychology, Princeton University, Princeton, NJ 08544; <sup>b</sup>Department of Psychology, Harvard University, Cambridge, MA 02138; and <sup>c</sup>White House Social and Behavioral Sciences Team, Washington, DC 20006

Edited by Uta Frith, University College London, London, United Kingdom, and approved October 30, 2015 (received for review June 17, 2015)

**How do people understand the minds of others? Existing psychological theories have suggested a number of dimensions that perceivers could use to make sense of others' internal mental states. However, it remains unclear which of these dimensions, if any, the brain spontaneously uses when we think about others. The present study used multivoxel pattern analysis (MVPA) of neuroimaging data to identify the primary organizing principles of social cognition. We derived four unique dimensions of mental state representation from existing psychological theories and used functional magnetic resonance imaging to test whether these dimensions organize the neural encoding of others' mental states. MVPA revealed that three such dimensions could predict neural patterns within the medial prefrontal and parietal cortices, temporoparietal junction, and anterior temporal lobes during social thought: rationality, social impact, and valence. These results suggest that these dimensions serve as organizing principles for our understanding of other people.**

social cognition | theory of mind | mentalizing | functional magnetic resonance imaging | multivoxel pattern analysis

The human mind plays host to a panoply of thoughts, feelings, intentions, and impressions. External observers can never directly perceive these mental states—one can never see “nostalgia” nor touch “awe.” Nevertheless, humans are quite adept at representing other people’s internal states. Our ability to perceive and distinguish among the rich set of others’ mental states serves as the bedrock of human social life. We understand the fine differences between pure joy and *schadenfreude* and judge a friend’s glee accordingly. Our ability to distinguish a partner’s sympathy from sarcasm can make a world of difference to a relationship. Legal decisions frequently hinge on nuanced mental distinctions such as that between inattention and intentional neglect. How do people navigate such complexities in others’ internal mental worlds?

One crucial tool for any navigator is a compass: a set of dimensions that help organize the contents of the world. By attending to the position of others’ mental states on key dimensions, humans might reduce the complexity of others’ minds to just a few essential elements—coordinates on a map. Might navigators of the world of mental states make use of such an intuitive compass? Research in other domains of cognition suggests such organization might be possible: The brain has a demonstrated capacity for extracting and capitalizing on useful regularities in the world. For example, our object representation system makes use of dimensions such as size and animacy to organize its processing tracts (1). Here, we explore the possibility that similar principles may organize our representations of other people’s minds.

Decades of research in social cognitive neuroscience, primarily using functional magnetic resonance imaging (fMRI), have already implicated a well-defined set of brain regions in the process of thinking about mental states: Thinking about the lives and minds of others reliably engages a network including the medial prefrontal cortex (MPFC), medial parietal cortex (MPC), temporoparietal junction (TPJ), superior temporal sulcus (STS), and the anterior temporal lobe (ATL) (for a review, see refs. 2 and 3). However, this relatively young field has yet to explain how the

social brain’s hardware processes the richness and complexity of others’ mental states. Fortunately, research in psychology supplies a set of theories regarding how people might organize their knowledge of mental states. The dimensions of these theories include valence and arousal (4, 5), warmth and competence (6, 7), agency and experience (8), emotion and reason, mind and body (9), social and nonsocial (2, 10, 11), and uniquely human and shared with animals (12). Any of these dimensions might plausibly play a role in organizing our understanding of mental states. But which, if any, do we spontaneously use during mentalizing? If a dimension actually matters to the way people typically think about others’ mental states, we should see evidence that the brain organizes its activity around that dimension. However, merely locating where in the brain mental state processing occurs—as social neuroscience has done so well already—cannot tell us how these regions represent mental states.

Fortunately, new analytic techniques in functional neuroimaging, under the umbrella of multivariate or multivoxel pattern analysis (MVPA), enable us to bridge these levels of analysis. MVPA examines activity in distributed sets of voxels, allowing for discrimination between stimuli by their associated patterns of activity even when absolute magnitudes of activity remain constant. In this study, we use the form of MVPA known as representational similarity analysis (13) to test which psychological dimensions organize people’s understanding of mental states. These analyses work by measuring the extent to which neural patterns of activity can be predicted from theories of representational organization. To illustrate, the dimension “arousal” would predict that “ecstasy” and “rage” are represented very similarly in the brain because both

## Significance

This study uses advanced functional neuroimaging analyses to test both existing and novel psychological theories about how we understand others' minds. Analyses show that three dimensions—rationality, social impact, and valence—account for almost half of the variation in the neural representation of mental states, the most comprehensive theory to date regarding our ability to think about others' minds. These findings both inform long-standing debates within social psychology about theory of mind and generate testable predictions about how our neural hardware supports our ability to mentalize.

Author contributions: D.I.T., M.A.T., J.M.C., and J.P.M. designed research; D.I.T., M.A.T., and J.M.C. performed research; D.I.T. and M.A.T. analyzed data; and D.I.T., M.A.T., and J.P.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The neuroimaging data have been deposited in the Harvard University Dataverse, <https://dataverse.harvard.edu> (accession ELLLZM).

See Commentary on page 19.

<sup>1</sup>D.I.T. and M.A.T. contributed equally to this work.

<sup>2</sup>To whom correspondence should be addressed. Email: [mthornton@fas.harvard.edu](mailto:mthornton@fas.harvard.edu).

This article contains supporting information online at [www.pnas.org/lookup/suppl/doi:10.1073/pnas.1511905112/-DCSupplemental](http://www.pnas.org/lookup/suppl/doi:10.1073/pnas.1511905112/-DCSupplemental).

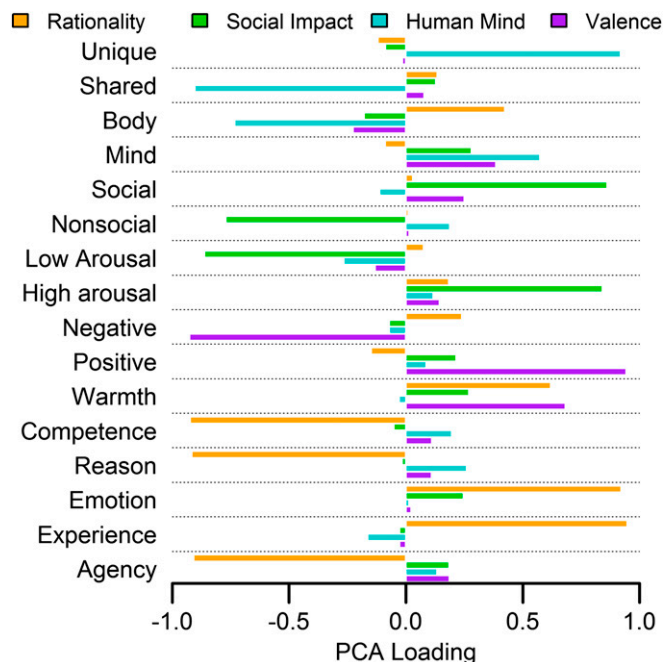
are similarly intense mental states. In contrast, the dimension “valence” would predict that “ecstasy” and “rage” are represented very differently in the brain because one state is very positive, whereas the other is very negative. Both predictions can be tested by measuring the extent to which patterns of neural activity elicited by thinking about a person in ecstasy are similar to those elicited by thinking about a person in a fit of rage. Each dimension makes thousands of predictions about the similarity of each mental state compared with each other mental state; representational similarity analysis allows us to assess the accuracy of all of these predictions simultaneously. Thus, we can test which psychological dimensions capture the way the brain encodes others’ mental states.

## Results

**Refining Psychological Theories.** We used 16 dimensions extracted from the psychological literature as a starting point for developing a theory of mental state representation: positive, negative, high arousal, low arousal, warmth, competence, agency, experience, emotion, reason, mind, body, social, nonsocial, shared, and unique. Note that these initial dimensions are nominal—in many cases they merely represent different poles of the same underlying variable—but we initially analyze them separately to remain maximally agnostic to the possible covariance between them. To determine what predictions each dimension would make about mental state representation—that is, which mental states were predicted to be similar or different with regard to each dimension—we used a large online sample ( $n = 1,205$ ) to measure the position of 166 mental states on each dimension. Ratings across many of the dimensions were highly correlated (Fig. S1). We distilled the overlapping intuitions embodied in the original dimensions down to a smaller set of nonredundant dimensions using principal component analysis (PCA).

The PCA revealed a much simpler set of four orthogonal dimensions, each with easily interpretable loadings (Fig. 1). The first component, which we term “rationality,” loaded highly in one direction on the original dimensions experience, emotion, and warmth, and loaded highly in the opposite direction on competence, reason, and agency. States such as embarrassment and ecstasy occupy one pole of this dimension whereas the other pole is occupied by states such as planning and decision. The second component, which we term “social impact,” loaded positively on the dimensions high arousal and social, and negatively on low arousal and nonsocial. States such as dominance, friendliness, and lust rate highly on social impact whereas sleepiness and pensiveness rate as minimally impactful. The third component, which we term “human mind,” loaded positively on unique to humans and mind, and negatively on shared with other animals and body. States high in human mind include those like imagination or self-pity whereas states such as fatigue and stupor are considered more physical in nature. The fourth component, which we term “valence,” loaded positively on positive and warmth, and negatively on negative. Positive states include affection and satisfaction whereas negative states include disgust and disarray. From each PCA dimension, we derived predictions about the similarity of each mental state to the others by calculating their psychological similarity as the absolute difference between the positions of mental states on each dimension. These predictions were tested against the neural data using representational similarity analysis, allowing us to see whether patterns of neural activity elicited by thinking about mental states reflected each dimension.

**Neural Patterns Representing PCA Dimensions.** Participants were scanned while performing a task designed to elicit their thoughts about 60 mental states (Table S1). On each trial, participants saw the name of a mental state (e.g., “awe”) and decided which of two scenarios would better evoke that mental state in another person (e.g., “seeing the pyramids” or “watching a meteor shower”). This task allowed us to estimate neural representations for each of 60 mental states by averaging the patterns elicited across the varied scenarios. We estimated the pairwise similarity of the neural representations of the 60 states by correlating their activity patterns.

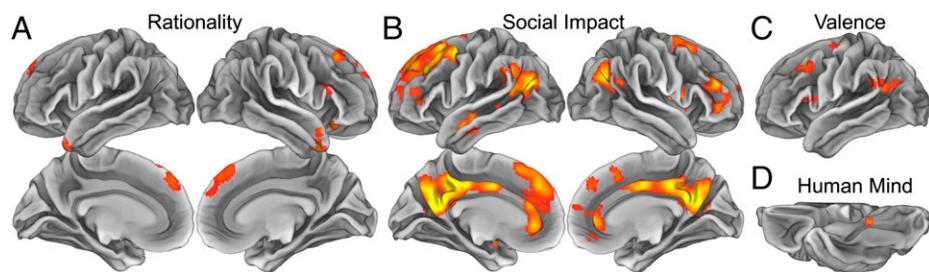


**Fig. 1.** Principal component loadings. Principal component loadings of the 16 existing theoretical dimensions onto the optimal four-dimensional solution.

These measures of neural similarity were then regressed onto the predictions of psychological similarity made by the four PCA-derived dimensions. For example, if mental states that rated similarly on the valence dimension (such as “affection” and “inspiration”) also elicited similar neural patterns of activity, the regression would reveal that valence was a strong predictor of neural pattern similarity. We would take this result as evidence that mental state representations—embodied by these neural patterns—were indeed organized by valence. This process was conducted repeatedly using local patterns extracted from throughout the brain of each participant. Regression maps for each dimension were combined across participants using  $t$  tests, thus revealing which dimensions reliably organized mental state representations in each region of the brain.

This analysis revealed that three PCA-derived psychological dimensions organize the way the brain represents mental states. Most regions implicated in mental state representation fell within a network of regions previously implicated in social cognition (Fig. 2 and Table S2). The “rationality” dimension predicted the similarity of patterns of neural activity in portions of the dorsolateral prefrontal cortex (DLPFC), ventral lateral prefrontal cortex (VLPFC), dorsal medial prefrontal cortex (DMPFC), lateral orbitofrontal cortex (OFC), and anterior temporal lobe (ATL) bilaterally (Fig. 2A). The “social impact” dimension robustly predicted neural pattern similarity in a widespread set of regions, including significant clusters in the DLPFC, VLPFC, DMPFC, VMPFC, anterior cingulate cortex (ACC), posterior cingulate cortex (PCC), precuneus, temporoparietal junction (TPJ) extending into the posterior superior temporal sulcus (pSTS) and ATL (Fig. 2B). The valence dimension predicted neural pattern similarity in a completely left-lateralized set of regions including the DLPFC, VLPFC, and TPJ (Fig. 2C). Finally, the “human mind” dimension captured a spatially restricted set of neural patterns, predicting representations in only a single region in the posterior parahippocampal cortex (Fig. 2D).

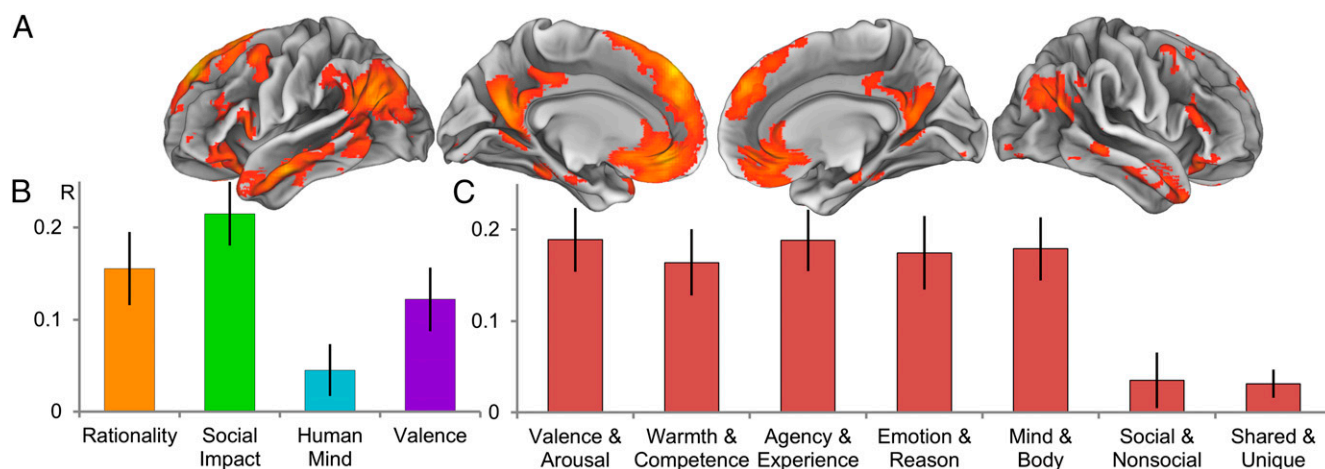
This analysis identified regions of the brain within which local patterns of activity were predicted by the PCA-based models. To test whether relevant patterns of activity were represented in a more distributed manner, we conducted a network-wide analysis. In this analysis, we extracted a single set of activity patterns from across the entirety of a neural network sensitive to mental state



**Fig. 2.** Searchlight results indicating regions sensitive to the (A) rationality, (B) social impact, (C) valence, and (D) human mind of others' mental states. Within the yellow/orange regions, the similarity of patterns elicited by thinking about mental states can be explained in terms of the corresponding social cognitive dimension extracted from existing theories via PCA ( $P < 0.05$ , corrected). Representational similarity searchlight analyses were conducted on each participant and combined through one-sample random-effects  $t$  tests.

content. As with the whole brain analysis, the neural similarity of each pair of mental states was estimated, and the results were correlated with the predictions of the PCA-derived dimensions. Results showed that three dimensions significantly predicted network-level patterns: rationality [ $r = 0.16$ ; 95% bootstrap confidence interval (CI) (0.06, 0.20)], social impact [ $r = 0.21$ ; 95% bootstrap CI (0.12, 0.26)], and valence [ $r = 0.12$ ; 95% bootstrap CI (0.04, 0.17)]. The human mind dimension [ $r = 0.05$ ; 95% bootstrap CI (-0.01, 0.10)] did not (Fig. 3B). Results of a multidimensional scaling analysis (Fig. S2) allowed us to estimate that the dimensions of rationality, social impact, and valence collectively account for approximately one-third of the variance in neural patterns underlying mental state representation (weighted total  $R^2 = 0.33$ ) (SI Text). Disattenuating this value by dividing it by the reliability of the neural similarity ( $\alpha = 0.69$ ) yielded a final  $R^2 = 0.47$ . The results of the network analyses were highly robust to different analytic approaches (SI Text). Statistically controlling for the influence of scenario concreteness, complexity, and familiarity did not produce any qualitative changes in the outcomes. Using independent component analysis (ICA) instead of PCA to generate dimensions, conducting the analysis with Spearman rank correlations, and using a meta-analysis-based feature selection method all produced very similar results. Further, results were not contingent on the use of statistical significance: The same three dimensions emerged from a model selection technique based on cross-validation performance (14) (Fig. S3). Finally, allowing two-way interactions between dimensions did not alter the significance of the main effects although three significant interactions were observed: human mind with rationality, human mind with social impact, and social impact with valence.

**Neural Patterns Representing Theoretical Models.** Although the primary purpose of this study was to discover the organization of mental state representation, we also tested whether the seven psychological theories from which we drew our PCA dimensions could predict neural representations of mental states. To do so, we repeated the whole brain and network-level representational similarity analysis with the original psychological dimensions. Whole brain analyses on each of the seven extant theoretical models revealed regions of the brain within which patterns of neural activity were predicted by each model (Fig. 4 and Table S3). The valence and arousal model (Fig. 4A) predicted patterns of activity in a number of regions, including the PCC, ACC, bilateral lateral temporoparietal cortex, left lateral and anterior temporal cortex, bilateral DLPFC, and both rostral and caudal portions of the DMPFC. The warmth and competence model (Fig. 4B) predicted patterns of activity in the left TPJ, rostral and caudal DMPFC, bilateral ATL, bilateral VLPFC, and bilateral DLPFC. Agency and experience (Fig. 4C) and emotion reason (Fig. 4D) produced very similar results, an unsurprising outcome given the degree of correlation between these models. These models both predicted patterns of activity in the VMPFC, rostral DMPFC, bilateral ATL, bilateral VLPFC and DLPFC, and portions of the lateral temporal cortex. The mind and body dimensions (Fig. 4E) predicted patterns in a proximal but distinct set of regions to those discussed above, including the ACC, PCC, TPJ, and portions of the lateral prefrontal cortex. Sociality (Fig. 4F) and human uniqueness (Fig. 4G) models both predicted much less extensive clusters of activity, with both appearing in the precuneus and uniqueness also appearing in a posterior portion of the parahippocampal gyrus.



**Fig. 3.** Network-wide representational similarity analysis. (A) Whole brain ANOVA used for feature selection (voxelwise  $P < 0.0001$ ). Different mental states reliably elicited different levels of univariate activity within these regions. (B) Bar graphs of model fits for dimensions derived via principal component analysis from existing psychological theories. (C) Bar graphs of model fits for existing psychological models. All model fits are given in terms of Pearson product-moment correlations between neural pattern similarity and model predictions, with error bars indicating bootstrapped SEs. Note that bars in B refer to individual dimensions derived via PCA whereas bars in C indicate the performance of full multidimensional theories. The theoretical advantage of the synthetic model presented here can thus be seen by comparing any one bar in C with the combination of the three significant bars in B.

Finally, we tested the degree to which each of the seven theoretical models predicted patterns of neural activity in a distributed manner. At the network level, the predictions of five of seven theoretical models were significantly correlated with neural similarity (Fig. 3C)—valence and arousal [ $r = 0.19$ , 95% bootstrap CI (0.10, 0.23)], warmth and competence [ $r = 0.16$ , 95% bootstrap CI (0.07, 0.21)], agency and experience [ $r = 0.19$ , 95% bootstrap CI (0.09, 0.22)], emotion and reason [ $r = 0.17$ , 95% bootstrap CI (0.06, 0.22)], and mind and body [ $r = 0.18$ , 95% bootstrap CI (0.09, 0.22)]—all with statistically indistinguishable effect sizes. Two theoretical models did not predict network level patterns: social vs. nonsocial [ $r = 0.04$ , 95% bootstrap CI (−0.03, 0.09)] and shared vs. unique [ $r = 0.03$ , 95% bootstrap CI (−0.003, 0.06)].

## Discussion

The current study used fMRI and representational similarity analysis to explore the dimensions that organize our representations of other people's internal mental states. We used dimensions from the existing psychological literature on mental states as a springboard for generating four nonredundant, easily interpretable dimensions and tested which dimensions organize patterns of neural activity elicited by considering others' mental states. Results indicated that neural activity patterns within the network of regions sensitive to others' mental states are attuned to three dimensions: rationality, social impact, and valence. These dimensions account for nearly half of the variation in the neural representation of mental states, constituting the most comprehensive theory to date regarding how we understand others' minds.

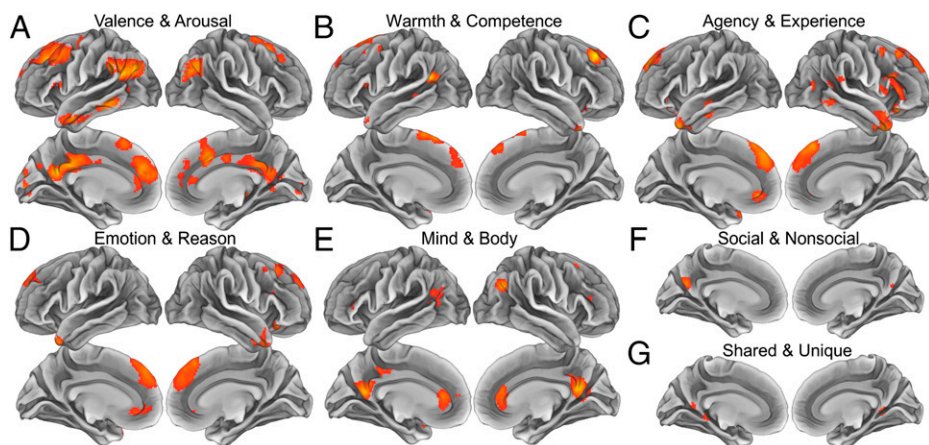
What significance do these three dimensions hold? One of these dimensions, termed "rationality," has arisen across disparate philosophical and psychological traditions. Here, it derives from theories in the domain of social cognition, including primarily experience and agency (8), warmth vs. competence (6, 7), and emotion vs. reason, an idea extending back at least as far as Plato. This dimension may also closely mirror theories outside the social domain, such as active vs. passive (15), system I vs. system II (16), and reflective vs. reflexive (17). The ubiquity of this distinction hints that it may reflect a deep principle of cognition. The results of the present study align with previous MVPA work (18) in suggesting that the brain spontaneously attunes to others' rationality. Knowing whether a person is experiencing a rational state or not may be particularly useful for certain social calculations. For example, it seems plausible that rationality assessments may help guide our decisions about whether people are responsible for their actions. These decisions in turn would shape the degree to which we take those actions into account during impression formation. These functions have been repeatedly associated with the DMPFC, one of the regions implicated in representing rationality (19–22).

A second dimension, termed "social impact," combines two well-known concepts: arousal and sociality. Social impact is the

most widely represented of the three dimensions identified here, suggesting that it may serve as a crucial ingredient in many different social computations. We did not anticipate the degree of covariation that these constructs displayed although this shared variation across seemingly disparate dimensions is clearly important, because sociality alone explains little neural pattern similarity. Validating and exploring the nature of this construct should be a topic for future research. Here, we suggest one possible explanation: A key property of another's mental state is how much that state is likely to affect one's self. For example, intense (i.e., high arousal) states are often more impactful than more moderate states. However, another person's rage, although highly arousing for them, may hold import for us only to the extent that it is directed outward at other people (i.e., social) rather than inward. Similarly, another's envy, although highly social, may hold import for us only in proportion to its intensity. Thus, whereas others' mental states might affect the self for many reasons, highly intense and social states may be most likely to do so.

The third dimension to emerge from this study, "valence," captures the difference between positive and negative mental states. This concept has long been implicated in social and affective processing (5). As such, it may come as no surprise that valence plays an important role in the organization of mental state representations. Of note, however, is that we find a unique spatial distribution associated with this dimension. Previous work has associated the processing of positive vs. negative stimuli with specific neural networks, including the mesolimbic dopamine system (23), as well as other limbic structures, such as the amygdala (24). Supplementary univariate analyses do show that the VMPFC, a region involved in reward and value more generally, tracks the positivity of mental states (Fig. S4). However, our MVPA results did not identify these regions but instead implicated left-lateralized cortical regions in the lateral prefrontal cortex and the angular gyrus. One possible explanation is that language supports the processing of mental state valence, but not other types of valence, a hypothesis here only preliminarily supported by the lateralization and the proximity of the valence regions to known language areas.

Together, the three significant dimensions described above explain nearly half of the reliable variance in the neural representation of mental states. While much remains unexplained (Fig. S5), this result appears quite promising. The social impact dimension alone predicts more variance than any of the original theoretical models; the combination of the three significant PCA-derived dimensions explains approximately twice the variance of the circumplex model, the most successful of the original theories. At the same time, given their significance to psychological theory, it is both reassuring and unsurprising that five of the seven original theories significantly predict neural pattern similarity. Notably, even theories that were originally geared toward explaining traits or groups, such as the stereotype content model, demonstrate their efficacy in the mental state domain.



**Fig. 4.** Searchlight results indicating the spatial distribution of mental state representations consistent with (A) the circumplex model of affect, (B) the stereotype content model, (C) the agency and experience model of mind perception, (D) emotion and reason, (E) mind and body, (F) social and nonsocial, and (G) shared with other animals and uniquely human. The similarity of patterns within the yellow/orange regions can be explained by their proximity to each other on the dimensions of the corresponding social cognitive models ( $P < 0.05$  corrected). Searchlight analyses were conducted on each participant and combined through one-sample random-effects *t* tests.

This finding raises the interesting possibility that the same dimensions organize neural activity about different types of social constructs.

In addition to informing us about the psychological question of interest—the organization of mental states—the current results also hint at the neural encoding scheme within the social brain network. By assessing the representation of mental states at two different levels of spatial organization—local activity patterns within spherical searchlights and broader activity patterns across the social brain network—the current study is well placed to bear on this issue. The results of the present study support the hypothesis that information is encoded by patterns of activity within localized brain regions, rather than across different regions. If local patterns did not encode social information but coarse patterns across the network did, the searchlight analysis would fail to produce results. Instead, we observe reliable encoding of mental state information in local patterns across the social brain, and explanatory power at the network level appears roughly in proportion to the cortical extent of their local encoding. As such, the current results provide no evidence that others' mental states are represented by interregional activity differences above and beyond the information already contained in local patterns. Interestingly, we find that two regions, the dmPFC and TPJ, each underlie multiple dimensions. Previous work has already heavily implicated these regions in mentalizing. The convergence of multiple dimensions on these nodes may help to explain their prominence in this domain.

Here, we have identified three dimensions that organize our representations of others' mental states. However, participants in this study thought about only the mental states of a nonspecific other. Do these same dimensions apply across different categories of "other"? For example, our understanding of a friend's happiness likely differs considerably from our concept of a stranger's happiness; our understanding of our own happiness likely differs considerably from others' happiness. Future work should endeavor to understand whether the dimensions we discovered here expand or contract in their importance on the basis of the person under consideration. We might expect such changes to be asymmetric across dimensions depending on one's relationship with the person experiencing the mental state. For instance, when considering a close friend's mental state, we might become more sensitive to valence differences but less sensitive to social impact (because all of the friend's states are more impactful).

We can also ask how these dimensions might apply across social cognition more generally. The current study used only lexical stimuli and tested these dimensions on only English-speaking adults. Do these dimensions apply to social cognition in other cultures? Do infants or other primates demonstrate any of the building blocks of these dimensions? Do these same dimensions apply when mentalizing about nonlinguistic content? Previous work on cross-modal emotion representation indeed suggests that visual and verbal emotional stimuli may be processed similarly (25, 26) although the full model has yet to be tested. We hope that the current data will provide a solid foundation for future research in these domains. It is also worth considering precisely which processes the imaging task taps. The task relies heavily on conceptual representations of mental states, and it is not entirely clear how strongly these concepts might guide other forms of mentalizing.

Finally, we should endeavor to ask why the social brain would organize its activity in accordance with the three dimensions discussed above and not others. The dimensions that shape mental state representations likely contribute to helping us solve problems in the social world. For example, we speculate that the three dimensions identified here might inform calculations regarding the threat posed by others: Valence could indicate the probability of help or harm; social impact would help estimate the likely magnitude of that that help or harm; and rationality would indicate the likely method of its expression (e.g., harm through a devious plot vs. an explosion of rage).

The present study derived four potential dimensions of mental state representation—rationality, social impact, human mind, and valence—from the existing psychological literature. We discovered

that three of these dimensions—rationality, social impact, and valence—predicted patterns of neural activity elicited across the social brain network by consideration of others' mental states. By discovering which dimensions the brain spontaneously uses to organize the domain of mental states, we have forged a deeper understanding of both human social cognition and its relationship to our own internal mental experience. These findings both inform long-standing debates within social psychology about theory of mind and can be used to generate novel predictions about how the brain supports our ability to mentalize.

## Materials and Methods

**Participants.** Participants ( $N = 20$ ) were recruited via the Harvard University Study Pool (16 female; mean age, 22.7 y; range, 18–27 y). A Monte Carlo simulation was used to determine participant and trial numbers consistent with adequate statistical power (*SI Text*). All participants were right-handed native speakers of English, reported no history of neurological problems, and had normal or corrected-to-normal vision. Participants provided informed consent in a manner approved by the Committee on the Use of Human Subjects at Harvard University.

**Experimental Design.** Participants underwent functional neuroimaging while considering another person experiencing a variety of mental states. The task elicited patterns of neural activity that reflect the representation of each state. On each trial, participants considered 1 of 60 mental states (*Table S1*). At the onset of the trial, one mental state term was presented for 1 s. This word remained on screen while two very brief scenarios associated with that mental state appeared for 3.75 s, one on the lower left side of the screen and one on the lower right side. Participants were instructed to report which of the two scenarios they thought would better evoke the mental state in another person. Participants indicated their response using a button box in their left hand by pressing either the middle finger for the left scenario or their index finger for the right scenario. There were no correct answers because both scenarios were pretested to elicit the scenario in question. Each trial was followed by a minimum 250-ms fixation and a randomized jittered fixation period (mean 1.67 s, range 0–10 s, in 2.5-s increments). During scanning, participants saw each of the 60 mental states on 16 occasions. Each state was presented once per run over the course of 16 consecutive runs of 405 s each. Participants judged a unique pair of scenarios on each trial; each of 16 scenarios was used only twice over the course of the experiment. Stimuli were presented with PsychoPy (27).

The 60 mental states in this study were selected to maximize observable differences based on survey ratings from a separate set of participants ( $n = 1,205$ ) (*SI Text*). Many of the theories under consideration made similar predictions about mental state representations. We pared down the information contained in the extant models using PCA. The PCA was conducted with respect to the 16 rating dimensions described above and the 60 mental states selected for the experiment. Varimax rotation was used to maximize the interpretability of the factors while maintaining their orthogonality (oblique rotation indicated that the orthogonal solution was satisfactory) (*SI Text*). Parallel analysis (28) and very simple structure (29) criteria were used to determine component number, with both indicating four factor solutions. The scenarios presented to subjects in this study were all written to be concise (fewer than five words), believable, devoid of personal pronouns, in the present tense, and maximally associated with their respective mental state. We selected an optimal set of scenarios using a genetic algorithm on survey ratings from a separate set of participants ( $n = 795$ ) (*SI Text*).

**Functional Imaging Procedure.** Functional data were acquired using a gradient-echo echo-planar pulse sequence with parallel imaging and prospective motion correction [repetition time, 2,500 ms; echo time (TE), 30 ms; flip angle, 90°] on a 3T Siemens Trio with standard 32-channel headcoil. Images were acquired using 43 axial, interleaved slices with a thickness of 2.5 mm and 2.51 × 2.51-mm in-plane resolution (field of view, 216 mm<sup>2</sup>; matrix size, 86 × 86 voxels; 162 measurements per run). Functional images were preprocessed and analyzed with SPM8 (Wellcome Department of Cognitive Neurology), using SPM8w. Data were first spatially realigned to correct for head movement and then normalized to a standard anatomical space (2-mm isotropic voxels) based on the ICBM 152 brain template (Montreal Neurological Institute).

A general linear model (GLM) was used to generate participant-specific patterns of activity for each mental state. The model included one regressor for each of the mental states, for a total of 60 regressors of interest. Events were modeled using a canonical hemodynamic response function and covariates of no interest (temporal and dispersion derivatives, session mean, run mean, linear trends, outlier time points, and six motion realignment

parameters). Boxcar regressors for events began at the onset of the presentation of the mental state. GLM analyses resulted in 60 *t*-value maps, one for each mental state, for each participant. In essence, these maps embody the average neural representation of each state.

We compared neural representations at each voxel in the brain using a searchlight procedure (30). Patterns of activity for each of the 60 mental states were extracted from participant's GLM-derived *t*-value maps using a spherical searchlight with 4-voxel radius (~9 mm). To compare the similarity of activity patterns for different mental states, we computed the Pearson correlation between each pair of patterns. Thus, two mental states that elicited highly correlated patterns of activity across the searchlight were considered to be more similar to each other. This searchlight procedure resulted in neural similarity matrices at each point in the brain: 60 × 60 matrices whose elements correspond to the correlations between the patterns of neural activity within that searchlight.

We used these estimates of neural similarity to test whether mental states were represented in a manner predicted by the four PCA-derived dimensions. To do so, we made similarity predictions for each dimension with respect to each pair of mental states by taking the absolute difference in their scores on the dimension in question. Multiple regression was used to determine how well the predictions of the PCA-derived dimensions accounted for neural similarity. These regressions generated four maps of unstandardized regression coefficients for each participant, one for each component. The participant-specific maps were smoothed (Gaussian 6-mm FWHM kernel) and entered into random effects analysis using one-sample *t* tests. The four resulting *t*-value maps indicate regions of the brain in which differences between mental states along each component. Results were corrected for multiple comparisons via a Monte Carlo simulation using the AFNI (31) 3dClustSim script (estimates of actual smoothness obtained from the four PCA maps and averaged; whole brain mask from the contrasts constrained voxel number). This simulation indicated that, with an uncorrected threshold  $P < 0.001$ , a 76-voxel extent was sufficient to yield a corrected threshold of  $P < 0.05$ . For visualization, statistical maps were rendered on the cortical surface using Connectome Workbench (32).

To test whether relevant patterns of activity were represented in a more distributed manner, we conducted an additional network-wide similarity analysis. In this analysis, we generated a single neural similarity matrix per

participant based on the pattern of activity across an independently defined network of neural regions. This network was defined using a whole brain omnibus repeated-measures ANOVA across the 60 mental states and 20 participants, which selected any voxels that showed different levels of activity across mental states (Fig. 3A). Due to the sensitivity of this analysis, voxels were selected at a conservative voxelwise threshold of  $P < 0.0001$ . The univariate nature of this approach appeared adequate as similar regions emerge from split-half searchlight reliability (Fig. S6). Note that, whereas this feature selection relied on the same data subjected to MPVA, it was independent of any of the dimensions being tested and thus did not yield biased results. Indeed, the network analysis based on these voxels produced results nearly indistinguishable from the same analysis conducted using voxels selected via a metaanalysis of mentalizing studies (S1 Text).

As with the searchlight analysis, in the network analysis, patterns of neural activity were extracted from the entirety of the feature selected area for each of the 60 mental states. These patterns were correlated to produce a single neural similarity matrix for each participant. These matrices were then averaged to produce a single group-level matrix. The group neural similarity matrix was Pearson-correlated with the similarity matrices generated from each of the four latent dimensions. To generate confidence intervals for these correlations, this procedure was repeated 10,000 times with group similarity matrices based on bootstrapped samples of the 20 participants.

We conducted analogous searchlight and network similarity analyses to test the seven theoretical models. The similarity between pairs of mental states was calculated as the (opposite of the) distance between the two mental states in the Euclidean space determined by the dimensions of each theory. This analysis diverged from that used for the PCA-based models only in that each theoretical model's predictions were independently correlated with neural similarity. This divergence was due to the substantial collinearity between the models, which was absent from the PCA-based models.

**ACKNOWLEDGMENTS.** We thank Talia Konkle, Brenda Li, Radhika Rastogi, Eve Wesson, and Ava Zhang. D.I.T. was supported by NIH Blueprint for Neuroscience Research Training Grant T90DA022759. M.A.T. and J.M.C. were supported by Graduate Research Fellowships from the National Science Foundation (DGE 1144152). M.A.T. was also supported by The Sackler Scholar Programme in Psychobiology. The views expressed in this article do not necessarily reflect the views of the General Services Administration or the United States Government.

- Konkle T, Caramazza A (2013) Tripartite organization of the ventral stream by animacy and object size. *J Neurosci* 33(25):10235–10242.
- Mitchell JP (2008) Contributions of functional neuroimaging to the study of social cognition. *Curr Dir Psychol Sci* 17(2):142–146.
- Van Overwalle F, Baetens K (2009) Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *Neuroimage* 48(3):564–584.
- Posner J, Russell JA, Peterson BS (2005) The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Dev Psychopathol* 17(3):715–734.
- Russell JA (1980) A circumplex model of affect. *J Pers Soc Psychol* 39(6):1161–1178.
- Cuddy AJ, Fiske ST, Glick P (2008) Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Adv Exp Soc Psychol* 40:61–149.
- Fiske ST, Cuddy AJ, Glick P, Xu J (2002) A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *J Pers Soc Psychol* 82(6):878–902.
- Gray HM, Gray K, Wegner DM (2007) Dimensions of mind perception. *Science* 315(5812):619.
- Forstmann M, Burgmer P (2015) Adults are intuitive mind-body dualists. *J Exp Psychol Gen* 144(1):222–235.
- Mitchell JP (2009) Social psychology as a natural kind. *Trends Cogn Sci* 13(6):246–251.
- Britton JC, et al. (2006) Neural correlates of social and nonsocial emotions: An fMRI study. *Neuroimage* 31(1):397–409.
- Haslam N (2006) Dehumanization: An integrative review. *Pers Soc Psychol Rev* 10(3):252–264.
- Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis: Connecting the branches of systems neuroscience. *Front Syst Neurosci* 2:4.
- Khaligh-Razavi SM, Kriegeskorte N (2014) Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Comput Biol* 10(11):e1003915.
- Osgood CE, Suci GJ, Tannenbaum PH (1957) *The Measurement of Meaning* (Univ of Illinois Press, Oxford, England).
- Kahneman D (2003) Maps of bounded rationality: Psychology for behavioral economics. *Am Econ Rev* 93(5):1449–1475.
- Heckhausen H, Gollwitzer PM (1987) Thought contents and cognitive functioning in motivational versus volitional states of mind. *Motiv Emot* 11(2):101–120.
- Corradi-Dell'Acqua C, Hofstetter C, Vuilleumier P (2014) Cognitive and affective theory of mind share the same local patterns of activity in posterior temporal but not medial prefrontal cortex. *Soc Cogn Affect Neurosci* 9(8):1175–1184.
- Mende-Siedlecki P, Cai Y, Todorov A (2013) The neural dynamics of updating person impressions. *Soc Cogn Affect Neurosci* 8(6):623–631.
- Mitchell JP, Cloutier J, Banaji MR, Macrae CN (2006) Medial prefrontal dissociations during processing of trait diagnostic and nondiagnostic person information. *Soc Cogn Affect Neurosci* 1(1):49–55.
- Mitchell JP, Neil Macrae C, Banaji MR (2005) Forming impressions of people versus inanimate objects: Social-cognitive processing in the medial prefrontal cortex. *Neuroimage* 26(1):251–257.
- Schiller D, Freeman JB, Mitchell JP, Uleman JS, Phelps EA (2009) A neural mechanism of first impressions. *Nat Neurosci* 12(4):508–514.
- Sabatinelli D, Bradley MM, Lang PJ, Costa VD, Versace F (2007) Pleasure rather than salience activates human nucleus accumbens and medial prefrontal cortex. *J Neurophysiol* 98(3):1374–1379.
- Garavan H, Pendergrass JC, Ross TJ, Stein EA, Risinger RC (2001) Amygdala response to both positively and negatively valenced stimuli. *Neuroreport* 12(12):2779–2783.
- Peelen MV, Atkinson AP, Vuilleumier P (2010) Supramodal representations of perceived emotions in the human brain. *J Neurosci* 30(30):10127–10134.
- Skerry AE, Saxe R (2014) A common neural code for perceived and inferred emotion. *J Neurosci* 34(48):15997–16008.
- Peirce JW (2007) PsychoPy: Psychophysics software in Python. *J Neurosci Methods* 162(1–2):8–13.
- Horn JL (1965) A rationale and test for the number of factors in factor analysis. *Psychometrika* 30(2):179–185.
- Revelle W, Rocklin T (1979) Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behav Res* 14(4):403–414.
- Kriegeskorte N, Goebel R, Bandettini P (2006) Information-based functional brain mapping. *Proc Natl Acad Sci USA* 103(10):3863–3868.
- Cox RW (1996) AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Comput Biomed Res* 29(3):162–173.
- Marcus DS, et al. (2011) Informatics and data mining tools and strategies for the human connectome project. *Front Neuroinform* 5:4.
- Brybaert M, Warriner AB, Kuperman V (2014) Concreteness ratings for 40 thousand generally known English word lemmas. *Behav Res Methods* 46(3):904–911.
- Kuperman V, Stadthagen-Gonzalez H, Brybaert M (2012) Age-of-acquisition ratings for 30,000 English words. *Behav Res Methods* 44(4):978–990.