HHS Public Access

Author manuscript

Cortex. Author manuscript; available in PMC 2021 April 01.

Published in final edited form as:

Cortex. 2020 April ; 125: 44–59. doi:10.1016/j.cortex.2019.12.012.

People represent mental states in terms of rationality, social impact, and valence: Validating the 3d Mind Model

Mark A. Thornton[*], **Diana I. Tamir**

Department of Psychology and Princeton Neuroscience Institute, Princeton University, Princeton NJ 08540, USA

Abstract

Humans can experience a wide variety of different thoughts and feelings in the course of everyday life. To successfully navigate the social world, people need to perceive, understand, and predict others' mental states. Previous research suggests that people use three dimensions to represent mental states: rationality, social impact, and valence. This 3d Mind Model allows people to efficiently "see" the state of another person's mind by considering whether that state is rational or emotional, more or less socially impactful, and positive or negative. In the current investigation, we validate this model using neural, behavioral, and linguistic evidence. First, we examine the robustness of the 3d Mind Model by conducting a mega-analysis of four fMRI studies in which participants considered others' mental states. We find evidence that rationality, social impact, and valence each contribute to explaining the neural representation of mental states. Second, we test whether the 3d Mind Model offers the optimal combination of dimensions for describing neural representations of mental state. Results reveal that the 3d Mind Model achieve the best performance among a large set of candidate dimensions. Indeed, it offers a highly explanatory account of mental state representation, explaining over 80% of reliable neural variance. Finally, we demonstrate that all three dimensions of the model likewise capture convergent behavioral and linguistic measures of mental state representation. Together, these findings provide strong support for the 3d Mind Model, indicating that is it is a robust and generalizable account of how people think about mental states.

1. Introduction

Humans enjoy rich mental lives, abuzz with a wide variety of different thoughts, feelings, perceptions, and intentions. These mental states exert influence over the behaviors that people are likely to perform (Frijda, 2004) and the states people are likely to experience in the future (Thornton & Tamir, 2017). Mental states play a central part in human social life. Social perceivers must understand their boss's happiness, their lover's anger, and taxi driver's alertness in order to successfully navigate a day in our highly social world. However, the complexity of the mind makes understanding others' mental states a serious

[*]Address correspondence to: Mark A. Thornton, mark.allen.thornton@gmail.com.

challenge. How can an observer ever fully understand the blinding throb of a headache, the sweeping exhilaration of a triumph, or the shattering chasm of grief? Trapped as we are within our own minds, it is doubtful we can ever fully grasp the richness of others' experiences or perfectly imprint the richness of our own states on others. However, our minds can still construct useful representations of the minds of others. These representations can discard some of the nuance and richness of others' experience as long as they retain the most essential elements necessary for navigating social life. The present investigation aims to characterize those essential elements – the principles people use to understand mental states.

Here we distill the complexity of mental states into a tractable form by placing them on a low-dimensional map. Maps of the physical world allow us to define locations in an efficient, useful form using just two dimensions: latitude and longitude. A mental map of the social world allows us to pinpoint the location of a mental state on relevant psychological dimensions. What are the latitude and longitude of mental state space? Prior psychological research offers many candidate dimensions. The circumplex model of affect suggests that the dimensions of valence and arousal determine the way people think about emotions (Posner, Russell, & Peterson, 2005; Russell, 1980). There is a long history of distinguishing between emotional and rational states, starting with Plato's chariot allegory (Hackforth, 1952), a distinction which many modern theories likewise retain (Heckhausen & Gollwitzer, 1987; Kahneman, 2003; Osgood, Suci, & Tannenbaum, 1957). More recently, researchers have suggested other potential dimensions of mental state representation, such as the sociality of a state (Britton et al., 2006).

These theories offer a starting point for synthesizing a comprehensive yet tractable model of mental state representation. In a recent investigation, we surveyed the literature and identified a total of 16 theory-derived dimensions of mental state representation (Tamir, Thornton, Contreras, & Mitchell, 2016). We then tested which dimensions could successfully capture how people's brains actually represents mental states. We arrived at the conclusion that three dimensions shape mental state representation: rationality, valence, and social impact. We refer to this set of three dimensions as the "3d Mind Model" from here on.

The first dimension of the model – 'rationality' – reflects whether a state is more cognitive or more emotional. At one pole are states high on agency, competence, and reason, such as planning and calculating; at the other pole are states high on emotion, experience, and warmth, such as ecstasy and outrage. We termed this dimension rationality to reflect one defining pole, with emotionality at the opposite pole. Although modern psychological and neuroscientific research questions whether there is a genuine distinction between cognition and emotion (Pessoa, 2008), this distinction remains popular in folk psychology, likely accounting for its prominence here. Understanding the rationality of another person's mental state may help people anticipate whether they are likely to carry out actions deliberately or rashly. Rationality also qualitatively resembles the dimension of potency which has featured prominently in previous dimensional studies of emotion (Russell & Mehrabian, 1977; Smith & Ellsworth, 1985).

The second dimension of the model – 'Social impact' – reflects whether a state is high intensity and socially directed or not. This dimension reflects the combination of two correlated features, arousal and sociality, leading us to choose the term social impact. At one pole are highly arousing and social states, such as lust and dominance; at the other pole are low arousal and nonsocial states, such as drowsiness and fatigue. The dimension is dissociable from whether a state is uniquely human (Tamir et al., 2016). Instead, it reflects states that are most likely to affect social relationships. The more likely someone is to engage with you (sociality) and the more intense their mental state (arousal) the more likely they are to have a substantial impact upon you.

The third dimension of the 3d Mind Model – 'valence' – indicates the extent to which a state is positive or negative. At one pole are highly positive and warm states, such as happiness and affection; at the other pole are highly negative states, such as misery and shame. As such, we termed this dimension valence. This name aligns with one of the names used in the circumplex model of affect, a prominent theory of emotion (Russell, 1980), and resembles the dimension of 'evaluation' – which distinguishes good from bad – in Osgood's semantic differential theory (Osgood et al., 1957). The valence of a mental state may also reflect the history of rewards – or reward prediction errors – which a person has recently experienced (Eldar, Rutledge, Dolan, & Niv, 2016). Accurately perceiving the valence of another person's mental state may be particularly useful in determining whether they hold helpful or harmful intentions.

As these dimensions indicate, the 3d Mind Model should be considered a synthesis of successful prior theories, rather than a new account of mental state representation. It incorporated prior models of the social-affective domains using dimensionality reduction. For example, the dimensions of valence and social impact largely align with valence and arousal in the circumplex model (Russell, 1980); dimensions of valence and rationality largely align with warmth and competence in the stereotype content model (Cuddy, Fiske, & Glick, 2008; Fiske, Cuddy, Glick, & Xu, 2002). Though research has validated the 3d Mind Model as a useful way for social perceivers to represent mental states (Thornton & Tamir, 2017; Thornton, Weaverdyck, & Tamir, 2019), in the present investigation, we expand our goals from synthesis to comprehensiveness, examining nearly 60 candidate dimensions of mental state representation to determine whether the 3d Mind Model provides the best description of how the brain represents other minds.

The 3d Mind Model can explain more than just the static representation of mental states: it can also predict mental state dynamics. The social world often rewards people for accurately predicting how other people will transition from one state to the next. If you can predict others' states, you can tailor your own behavior accordingly to better achieve your goals. The 3d Mind Model explains how people make these predictions about others' future mental states (Thornton & Tamir, 2017), as follows: the model is founded on the premise that people make predictions about mental states by attending to their locations in the 3d Mind space. The closer two states are in the space, the more likely they are to transition from one to the next. For instance, joy and gratitude are close together in the space because they are both positive emotions, whereas a negative emotion like disgust is far away from both. Correspondingly, the transitional probability between joy and gratitude is higher than the

transitional probability between disgust and gratitude. Proximity thus plays a key role in mental state predictions, in three ways: Proximity predicts i) actual transition likelihoods, as assessed by experience sampling, ii) people's perceptions of transition likelihoods, as assessed by both behavioral ratings, and iii) neural similarity between states, as assessed by functional neuroimaging (Thornton & Tamir, 2017; Thornton, Weaverdyck, & Tamir, 2019). That is, brain automatically encodes the rationality, social impact, and valence of each state, and in doing so, represents current states in a way that resembles likely future states. Together, these findings suggest that people use the dimensions in the 3d Mind Model to make accurate predictions about others' state transitions.

A well-founded model of mental states has the potential to explain foundational psychological phenomena and ground important theories (Tamir & Thornton, 2018). In the present investigation we subject the 3d Mind Model to its most rigorous and extensive test to date, to determine whether it can be viewed as the definitive model of this phenomenon. This test consists of three parts. In Study 1, we examine the robustness of the 3d Mind Model. To do so, we combine four open fMRI datasets from three previous studies in which participants made judgements about mental states (Tamir et al., 2016; Thornton, Weaverdyck, Mildner, & Tamir, 2019; Thornton, Weaverdyck, & Tamir, 2019). We perform a mega-analysis (a meta-analysis with participant-level data) across these studies to test whether each dimension of the model is a significant predictor of neural activity. These datasets vary in the particular states considered, the modality of the stimuli, and the design of the task, as well as in incidental features such as the site, scanner, and participants employed. By combining them, we not only bring to bear a large amount of data, but also examine whether the 3d Mind Model can robustly explain neural activity across all of these variations. In this way, the first test of the 3d Mind Model assesses the model's necessity – that is, are rationality, social impact, and valence each needed to explain mental state representation?

Study 2 examines the sufficiency of the 3d Mind Model – that is, do these three dimensions together comprise a complete theory of mental state representation, or are further dimensions required? To address this issue, we consider a set of 58 candidate mental state dimensions. These new dimensions include everything from the temporal profile of a mental state (e.g., whether a state begins or ends abruptly), to the social appropriateness of the state, to the sensory modalities used to sense it (e.g., vision, audition, etc.). Using the aforementioned fMRI data, we select the best possible set of dimensions for explaining mental state representation We then compare the results to determine whether the 3d Mind Model is sufficient or incomplete.

Finally, Study 3 assesses the generalizability of the 3d Mind Model. Unlike many previous models of affect, the 3d Mind Model is unique in that its initial validation came from explaining patterns of brain activity, rather than behavior. Studies 1 and 2 extend this validation, demonstrating that the model can indeed account for much of the neural representation of mental states. A comprehensive model of mental state representation ought to explain this neural variance, but it also needs to explain other forms of mental state representations as well. In Study 3, we examine the ability of rationality, social impact, and valence to predict two convergent indices of mental state representation: behavioral

judgements about mental states and the natural language semantics of mental state words. If the 3d Mind Model generalizes to these measures, this would provide evidence of its applicability to wide-ranging psychological phenomena.

Together, these three studies will allow us to answer three key questions about the nature of mental state representation. First, do rationality, social impact, and valence explain the way the brain represents others' thoughts and feelings? Second, how close are these three dimensions to a complete account of mental state representation? Third, do these dimensions generalize to other representations of mental states, such as in behavior and semantics? By answering these questions, we aim to bring the field closer to a comprehensive understanding of how people make sense of each other's thoughts and feelings.

2. Study 1: Mega-analysis of mental state representation

In Study 1, we conducted a mega-analysis – a meta-analysis based on individual participant data (Sung et al., 2014) – to assess the robustness of the 3d Mind Model. Specifically, we tested whether the proximity between mental states on the dimensions of rationality, social impact, and valence predicted the similarity between neural representations of mental states. To this end we drew upon four existing fMRI datasets from three previous studies. We will refer to these four datasets as Studies 1A-D hereafter. Data and code for each of these studies has already been deposited on the Open Science Framework (OSF) and are freely available. Study 1A corresponds to Tamir, Thornton, Contreras, and Mitchell (2016), with data available at https://osf.io/3qn47/. Study 1B corresponds to Study 2 of Thornton, Weaverdyck, Mildner, and Tamir (2019), with data available at https://osf.io/hp5wc/. Study 1C corresponds to Thornton, Weaverdyck, and Tamir (2019), and is available at https://osf.io/x32te/. Study 1D corresponds to Study 1 of Thornton, Weaverdyck, Mildner, and Tamir (2019), and is available at https://osf.io/8n9ze/. The processed data from these studies necessary for the present investigation is available at https://osf.io/6zmnc/. This OSF repository also contains the data which was newly collected as part of the present investigation, as well as the analysis code specific to this paper. No part of Studies 1A-C (nor Studies 2 or 3) procedure or analyses were preregistered prior to the research being conducted. We report how we determined our sample sizes, all data inclusion/exclusion criteria, whether inclusion/exclusion criteria were established prior to data analysis, all manipulations, and all measures in the study. Exclusion criteria were set prior to conducting inferential analyses in all studies. In the interests of clarity and concision, we do not fully recapitulate the methods of each previous study here. Instead, we focus on the essential elements of these datasets, as they relate to the present investigation.

2.1 Materials and Methods

2.1.1 Participants—A total of 120 participants were recruited in Studies 1A-D. Of these, seven participants were excluded: four due to excessive head motion (Study 1B), one due to data loss in image reconstruction (Study 1C), and two due to low behavioral response rates (Study 1D). This left a final imaging sample size of 113 people (70 female, 42 male, 1 nonbinary; mean age = 21, age range = 18–31) across Studies 1A-D (see Table 1 for breakdown). Sample sizes were determined by power analyses specific to each of the

original studies. Participants were right-handed or ambidextrous, fluent in English, had normal or corrected-to-normal vision, reported no history of neurological problems, and were screened for standard MRI safety exclusion criteria. All participants provided informed consent. Study 1A received ethical approval from the Committee on the Use of Human Subjects at Harvard University. Studies 1B-D and all other data collections received ethical approval from the Institutional Review Board at Princeton University.

2.1.2 Experimental paradigms—Participants in Studies 1A-D completed mental state judgement tasks in the fMRI scanner. On each trial of these task, participants saw the name of a single mental state, such as awe. In Studies 1A, 1B and 1D, participants saw each mental state paired with two scenarios (e.g., “seeing the pyramids” or “watching a meteor shower” for awe). They then judged which of those two scenarios they thought would best elicit the given mental state. In Study 1A, participants made these judgements with respect to a generic other person. In Study 1B participants made these judgements about three different target people: the self, a close target (e.g., friend or family member), and a far target (a biography artificially constructed to be dissimilar and dislikeable to the participant). In Study 1C, participants again judged a generic other, but the task was slightly different: only one scenario was presented with each state, and participants rated how much it would elicit that state on a Likert scale. In Study 1D, participants again compared two scenarios, but these scenarios were presented in the form of images instead of text. Participants in Study 1D judged the states of the self and far targets, as in Study 1B. The lengths of trials, number of trials per run, and number of runs all varied across studies (Table 1).

The mental states presented to participants varied between studies. In Study 1A, 60 different mental states were selected from an exhaustive list of 166 state terms (Tamir et al., 2016). These states were chosen to uniformly cover the larger state space, while being minimally redundant with one another. Studies 1B-D all used subsets of the 60 states presented in Study 1A (Table 1). In Studies 1B and 1D, the states were chosen to maximize a complex objective function (Thornton, Weaverdyck, Mildner, et al., 2019). In Study 1C, the states were chosen to maximize pairwise asymmetries in perceived mental state transition probabilities (Thornton, Weaverdyck, & Tamir, 2019).

In addition to these differences in mental states, the scenarios paired with each state also varied across studies. Scenarios in Studies 1A-C were all drawn from 36 short written scenarios generated for each of the 60 states in Study 1A. Of these, 16 scenarios were selected for each state in Study 1A, 12 scenarios per state in Study 1B, and 20 scenarios per state in Study 1C. Selection was performed via a genetic algorithm which aimed to choose scenarios which maximally elicited each state, while also minimizing incidental differences (e.g., character length, difficulty) between scenarios across states (Tamir et al., 2016). In Study 1D, an initial set of 1234 image-based scenarios was reduced to 24 images per state by removing outliers based on visual features. A genetic algorithm similar to that applied to the text scenarios was then used to select final sets of 12 scenarios for each state.

Participants in each of the imaging experiments also provided responses on a number of individual difference surveys and questionnaires. The specific measures in question are described in the corresponding publications. These data were collected for use in separate

research on individual differences in mental state representation, and thus not analyzed as part of the prior investigations or this paper.

2.1.3 Imaging procedures—All imaging participants were scanned using Siemens brand 3 T scanners (Siemens, Erlangen, Germany) with standard 32 or 64 channel head coils. Functional images were obtained with wholebrain coverage and spatial resolution varying between 2 mm^3 and 2.5 mm^3 isometric voxels. The time to repetition (TR) range from 1400 ms to 2500 ms. Study 1A used parallel imaging. Studies 1B-D used simultaneously multi-slice imaging. High resolution anatomical images were also collected from participants in all studies for the purpose of normalization. All studies except for Study 1A also collected field maps for the purposes of correcting inhomogeneities. Table 2 reports detailed parameters for each study.

Preprocessing and general linear modeling (GLM) were conducted separately in each study, as reported in the corresponding publications. Study 1A was processed using SPM8 (Wellcome Department of Cognitive Neurology, London UK) with the SPM8w wrapper (https://github.com/ddwagner/SPM8w). Studies 1B-D were analyzed using a multi-package pipeline (https://github.com/PrincetonUniversity/prsonpipe). Corrections for slice timing and head motion were performed using FSL (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012). Coregistration and normalization were performed using SPM8's DARTEL (Ashburner, 2007). The GLM was implemented using SPM12 (Wellcome Department of Cognitive Neurology, London, UK) with the SPM12w wrapper (https://github.com/wagner-lab/spm12w). We retain these differences in processing as an additional generalizability challenge to our model. That is, if the effects of rationality, social impact, and valence are as general as we suggest, then they should be able to predict neural data regardless of incidental differences in image processing pipelines. All images were normalized to the ICBM 152 template (Montreal Neurological Institute). To preserve fine-grained patterns of activity, explicit smoothing was not applied to any of the fMRI data used in this investigation.

We conducted GLMs to measure patterns of brain activity associated with each mental state. Boxcar regressors were generated for each mental state based on the onsets and durations of each trial in the respective tasks. In Studies 1B and 1D, participants considered the mental states of multiple targets. In these studies, boxcar regressors were generated for each state, separately for each target. Boxcar regressors were convolved with a canonical hemodynamic response function. They were then entered into the GLM along with nuisance regressors including run means, run trends, and six-parameter head motion. Contrasts were computed for each mental state (or mental state by target) regressor against baseline, producing wholebrain patterns of regression coefficients (used in Studies 1B-D) and t-statistics (used in Study 1A) associated with each mental state.

We extracted patterns for each mental state within a network defined by its sensitivity to mental state content in Study 1A (Tamir et al., 2016). This network included 10,216 non-contiguous voxels concentrated in regions including ventral and dorsal medial prefrontal cortex., the superior temporal sulcus extending from the temporoparietal junction to the anterior temporal lobe, and the medial parietal lobe (Figure 1A). We vectorized the patterns extracted from these regions, and computed the Pearson correlation distance ($1 - r$) between

each pair of patterns to produce neural representational dissimilarity matrices – a matrix that reflects how similar and dissimilar each state is to each other state. We produced one such dissimilarity matrix for each participant in each of the four studies.

In addition to the primary mega-analysis – which was conducted within the social brain network as a whole – we also conducted a whole-brain analysis. This analysis complements the network analysis by providing information about the spatial distribution of social information in the brain. To this end we extracted dissimilarity matrices from each of 200 parcels distributed throughout the cortex (https://identifiers.org/neurovault.collection:2099). These parcels were defined by meta-analytic functional co-activation, assessed using Neurosynth (De La Vega, Chang, Banich, Wager, & Yarkoni, 2016). We used parcels instead of a searchlight approach to better manage the computational requirements of the mega-analysis.

2.1.4 Statistical analysis—The mega-analysis was designed to test whether the brain represents mental states using the dimensions of rationality, social impact, and valence. If, for example, rationality matters to the way that people process mental states, then two states that are highly rational should look similar in the brain; if instead one state is highly rational and one highly emotional, then those two states should look very different in the brain. In this analysis, we test whether neural patterns reflect each of the three dimensions in this way, first in each of Study 1A-D independently, and then in aggregate, as a mega-analysis across all four data collections.

First, we generated predictions about how the brain *should* represent each state if each of the dimension influenced it. In prior work, we used human ratings to determine the coordinates of each mental state on all three dimensions (Tamir et al., 2016). We used those coordinates to calculate the absolute distances between each pair of mental states, on each dimension. These distances served as our predictors in representational similarity analyses (Kriegeskorte, Mur, & Bandettini, 2008). Specifically, we regressed neural pattern dissimilarities – as described above (2.1.3) – onto these three dimension-based predictors for each participant in each study. This is a form of mixed (versus fixed) representational similarity analysis, because it fits separate parameters for each of the three predictors rather than assuming equal weights on each (Khaligh-Razavi, Henriksson, Kay, & Kriegeskorte, 2017). These parameter estimates – that is, ordinary least squares regression coefficients for each dimension – indicate the extent to which each psychological dimension's predictions are associated with neural dissimilarity, controlling for the other dimensions in the model. Although the rationality, social impact, and valence were designed to be orthogonal across the full set of mental states, each study used only a selection of states. Within these smaller sets of mental states, the three dimensions often deviated from this orthogonality, hence the need for multiple regression. To maintain a consistent scale across participants, each participant's dissimilarity estimates were z-scored prior to being entered into the regression. In the cases of Studies 1B and 1D, which featured multiple target people, the neural similarity estimates were first averaged across the targets and then z-scored.

We next tested the statistical significance of the regression coefficients for rationality, social impact, and valence within each study. To do so, we computed Cohen's d – a standardized

effect size measure – for each study. In this case we computed *d* for each dimension by taking the average regression coefficient for that dimension across participants within a study, and then dividing that average by the corresponding standard deviation. This was done to maintain a similar scale for the effects across the four studies. We then used percentile bootstrapping to compute a 95% confidence interval around each effect size, and thereby assess its statistical significance. This analysis allowed us to determine whether each dimension had a significant association with neural pattern dissimilarity in separately in each study.

Finally, we sought to assess the effects of rationality, social impact, and valence on neural activity across all four studies via a mega-analysis. We performed this analysis via a multi-level bootstrapping procedure which simultaneously resampled both studies, and participants within studies. On each iteration of this procedure, we first drew a random sample (with replacement) of the four fMRI datasets. Thus, on one iteration, we might get {Study 1B, Study 1C, Study 1C, and Study 1D}, and on another we might get {Study 1A, Study 1A, Study 1B, Study 1C}, and so forth. Next, within each of the studies drawn on a particular iteration, we would draw a random sample of that study's participants (again, with replacement). Importantly, if a given study was drawn more than once on a given iteration (e.g., two "copies" of Study 1A) then different random samples of participants would be drawn from each copy of that study. For each participant, we had already computed the regression coefficients for rationality, social impact, and valence. We used these coefficients to compute Cohen's *d*s for each bootstrapped study, just as we did when we analyzed each study separately. Finally, we averaged these *d*s across studies to produce a test-statistic (i.e., the mean *d*) for each dimension. This entire procedure was repeated 10,000 times to produce confidence intervals around the mean *d*s. By resampling at both participant- and study-levels, this analysis allows us to justify generalizable inferences not just to the population of participants, but also to the hypothetical population of similar studies.

We also conducted a whole-brain version of this analysis, by repeating the mega-analytic procedure in each of 200 parcels across the cortex. Separate regressions were conducted within each parcel in the same way as in the network-level mega-analysis. Significance testing was achieved via bootstrapping at the study level and sign-flipping permutation at the participant level (Nichols & Holmes, 2002). Multiple comparisons were controlled via the maximal statistic method with generalized pareto tail approximation (Winkler, Ridgway, Douaud, Nichols, & Smith, 2016).

2.2 Results

The mega-analysis analysis indicated that rationality, social impact, and valence each explained significant unique variance in the similarity between mental state-specific patterns of brain activity (Figure 1B). The largest effect was that of valence (mean $d = .89$, 95% CI = [.58, 1.25]) followed by social impact (mean $d = .61$, 95% CI = [.20, 1.10]), and rationality (mean $d = .43$, 95% CI = [.05, .93]). Across all four of the studies examined, we observed effects in the predicted (i.e., positive) direction for all three dimensions. The effect of valence was statistically significant in all four studies. The effect of social impact was significant in all but Study 1D. The effect of rationality was significant in Studies 1A and

1D. We also conducted an analysis to test the extent to which these effects varied across studies. This variance components analysis indicated that study-specific factors contributed 0–12% of the variance observed in 3d Mind Model effects (see Supplementary Information).

This mega-analysis provides evidence for the 3d Mind Model, within the social brain network. However, this analysis was restricted to regions previously implicated in mental state representation; it cannot tell us whether the 3d Mind Model is specific to these regions, or whether any regions outside of this network also contribute to mental state representation. To determine the spatial distribution of these effects across the cortex, we repeated the mega-analysis within each of 200 parcels distributed throughout the brain. The results (Figure 2) indicate that the effects of the 3d Mind Model are indeed largely confined to regions typically implicated in social cognition, including medial prefrontal and parietal cortices, and superior temporal sulcus extending from the anterior temporal lobe to the temporoparietal junction. The effects of social impact and valence were present throughout these regions, whereas the effect of rationality was present only in dorsal medial prefrontal cortex. Examination of uncorrected results from each study (Figure S1) indicates that the apparent spatial specificity of rationality may rather originate from greater inter-study heterogeneity, and smaller effect sizes.

2.3 Discussion

The results of a mega-analysis of four fMRI studies support our earlier findings that rationality, social impact, and valence shape neural representations of mental states (Tamir et al., 2016). Moreover, these effects appear to be robust to a large number of substantial and incidental factors, including the timing of the trials, the nature of the task, which particular mental states were considered, whose mental states were considered, which participants considered them, or what scanner those participants were imaged in. Overall, the results of Study 1 indicate that the 3d Mind Model can indeed reliably explain the way the brain encodes others' thoughts and feelings.

The results of the wholebrain parcel-based mega-analysis indicate rationality, social impact, and valence explain neural activity primarily within regions associated with social cognition. That is, the way that the brain encodes the 3d Mind Model is spatially specific. This suggests that these dimensions may be specific to social cognition, rather than reflecting domain general semantic properties. The wholebrain results also clarify the spatial scale at which the 3d Mind Model describes neural activity. Specifically, they indicate that rationality, social impact, and valence are reflected in local, fine-grained activity patterns, and not just interregional activity differentials.

3. Study 2: Optimal model selection and evaluation

In Study 1, a mega-analysis of fMRI studies indicated that rationality, social impact, and valence robustly predict the neural similarity between mental state representation. This analysis demonstrated that these dimensions are necessary for explaining how the brain representation mental states. However, there might well be other psychological dimensions which are just as necessary for explaining how the brain represents states. While all three dimensions significantly predicted brain activity, we do not yet know if they can explain all,

or even most, of the reliable mental state-related brain activity. If one imagines the 3d Mind Model as a building supported by three columns, it is not enough that each of those columns bears some weight: together those three columns must bear all of the weight of the roof, or the structure will collapse.

Study 2 tested the sufficiency of the 3d Mind Model. It did this in two ways: first, it examined an extremely broad set of different candidate dimensions (58 in total) to determine whether any combination of them explains neural representations of mental states better than the 3d Mind Model. Second, it compared the performance of the optimal model – whatever it may be – to an estimate of the best possible performance any model could hope to obtain. This procedure, known as a noise ceiling analysis, provided an indication of how well the 3d Mind Model (or whatever other model proves optimal) compares to a hypothetical perfect model of mental state representation.

3.1 Study 2A: Principal component analysis

3.1.1 Materials and Methods—The first step in assessing the sufficiency of the 3d Mind Model was to assemble and synthesize a large set of candidate dimensions. To this end, we used data from Study 1A in which a sample of 1,205 participants on Amazon Mechanical Turk rated the 60 states in that study on 16 theory-derived dimensions of mental state representation. To supplement these data, we collected new ratings from another sample of 1647 Mechanical Turk participants who rated the same 60 states on 42 additional dimensions. The full set of 58 candidate dimensions rated by participants are shown in Figure 3 and Table S1. We selected these additional dimensions to cover a broad range of features, such as the sensory modality used to perceive each state (e.g., to what extent can a given state be perceived from tone of voice?), the dynamic properties of each state (how abruptly does it begin or end? How long does it last? Can other states occur simultaneously?), and aspects of the context in states are expressed (e.g., how specific is it to a certain context, action, or type of person?). By including such a broad array of dimensions, we attempted to exhaust all reasonable possibilities for how the brain might explain mental states. We included every dimension we could think of or find in the literature to pit against or complement the 3d Mind Model.

The rated dimensions were not orthogonal; instead, many were highly correlated with each other. To reduce the dimensionality of this space down to a set of unique, non-redundant components, we performed principal component analysis (PCA). To identify the appropriate number of dimensions for compressing the data we used a technique known as bi-cross-validation (Owen & Perry, 2009). This procedure uses cross-validation to help ensure that the appropriate number of components is extracted regardless of which sample of dimensions or mental states happened to be considered (see Supplementary Information). Thus, the results of this analysis should generalize well to any sample of mental states and any sample of dimensions.

3.1.2 Results—Results of PCA bi-cross-validation indicated that optimal performance (i.e. minimal root mean square error [RMSE]) was obtained with 11 components (Figure S2). We thus computed a final PCA across all of the rating data using this number of

components, and a varimax rotation. Factor loadings from this final PCA with 11 dimensions and varimax rotation are shown in Figure 3 and Table S1.

To assess whether this PCA included the dimensions from the 3d Mind Model, we calculated the correlations between component scores from the PCA, and the rationality, social impact, and valence scores (Table 3) from our earlier research (Tamir et al., 2016). These results indicate that rationality maps onto PC 2 ($r = .97$), social impact maps onto PC 3 ($r = .90$), and valence maps onto PC 1 ($r = .96$).

3.2 Study 2B Model selection and evaluation

3.2.1 Materials and Methods—The results of the principal components analysis (3.1) provided 11 independent candidate dimensions that people might use to think about mental states. However, people likely do not spontaneously use all of these dimensions to think about mental states. Certain dimensions may be automatically called to mind whenever someone reasons about thoughts and feelings, whereas other dimensions may only be recalled or computed when they are specifically necessary. FMRI provides a passive measure of participants brain activity: data can be collected whether or not participants produce any externally observable behavior. This makes it a useful tool for measuring the way people spontaneously encode mental states. We used the four fMRI datasets described in Study 1 (Section 2) to determine which combination of dimensions reflects the optimal model for explaining neural representations of mental states. After model selection, we then evaluated the performance of the optimal model relative to an estimate of the maximal performance possible, given the reliability of the data (i.e., a noise ceiling).

Model selection needs to be conducted on different data than model evaluation to avoid overfitting, or overly optimistic estimates. To mitigate this possibility, we conducted selection and evaluation in separate halves of the data. Data were divided in half randomly by participant within each of Studies 1A-D. The entire model selection and evaluation procedure described below was repeated 10,000 times: 5000 random split halves, with each half being used for selection and evaluation in turn.

To select an optimal model, we first needed a measure of model performance. We used a modified version of the representational similarity analysis describe in 2.1.4 above. The neural pattern dissimilarity between mental states was regressed onto the distances between mental states on each of the dimensions in the model under evaluation (i.e., some combination of the 11 PCA dimensions). This regression was performed using non-negative least-squares regression (Khaligh-Razavi & Kriegeskorte, 2014). We evaluated the performance of this regression using cross-validation, within the half of data set aside for model selection. Specifically, we used a round-robin procedure, in which regressions were trained on each participant in one dataset (i.e., Studies 1A-C), and then tested on each participant in another dataset. Neural data from different target people (self, close, or far) in Studies 1B and 1D were modeled separately, but cross-validation was only performed across studies. The predictive performance of each regression was measured in terms of the Pearson correlation between predicted and observed neural dissimilarity. The round-robin continued until every dataset had been predicted based on every other dataset. Performance

was averaged across participants and datasets to yield a final statistic indicating how well a particular combination of dimensions could predict neural dissimilarity.

Having established a way to measure the performance of different models, we then sought to select the optimal combination of the 11 PCA dimensions. We achieved this using a greedy search procedure. This procedure started with an empty model, consisting of 0 of the 11 PCA dimensions. On each iteration, it computed how model performance would chance if each dimension was added to the model. It would then add the dimension which most improved performance. Thus, on the first iteration, it considered all 11 dimensions, and selected the dimension which would best improve the empty model (i.e., the single most predictive dimension). On the second iteration, it evaluated the 10 remaining dimensions, and selected the dimension which would most enhance performance when added to the first selected dimension, and so forth. The search procedure terminated when it discovered that there was no one remaining dimension which could be added to further improve model performance. The set of dimensions included at termination was considered the optimal model. Across these replicates, we counted how often each set of dimensions emerged as the best set during the model selection procedure. This allowed us to determine the modal optimal model – that is, the set of dimensions which were most often selected as optimal. In the process we also calculated how frequently each individual dimension was included in the optimal model.

Having established an optimal model, we next sought to evaluate its performance. This evaluation required two analyses. First, we estimated how well the optimal model would generalize to the held-out evaluation half of the data. As in model selection, this was achieved using non-negative least squares regressions with cross-validation in a round-robin across studies. However, in this instance, the model was fit to the model selection half of the data, and then evaluated in the other half. Second, we compared this measure of the performance of the optimal model to an estimate of the performance of the hypothetical ideal model. To do so, we performed a noise ceiling analysis (Kriegeskorte et al., 2008) in which we computed the maximal model performance possible given the reliability of the data. We estimated this noise ceiling by taking the average correlation between the model selection data and the held-out evaluation data, in the same round-robin used to evaluate optimal model performance. We calculated the proportion of maximal performance achieved by optimal model achieved by dividing the optimal model's performance by the noise ceiling. Results were averaged across replicates of the model selection and evaluation procedure.

3.2.2 Results—A model selection and evaluation procedures were conducted to determine the optimal model for explaining neural representations of mental states. This model pitted the 3d Mind Model against many other candidate dimensions (including supersets of its three dimensions) to determine which combination best explained neural pattern similarity.

First, we evaluated the results in terms of individual dimensions. The model selection procedure strongly supported the importance of rationality, social impact, and valence. PC 1, mapping on to valence, was included in 99.99% of all optimal models. PC 2, mapping on to

rationality, was included in 88.15% of all optimal models. PC 3, mapping on to social impact, was included in 80.37% of all optimal models. None of the other eight PCs rivaled these three: the next most prominent were PC 11, included in 24.22% of all optimal models, and PC 7, included in 22.74% of optimal models. All of the other PCs were included in fewer than 10% of optimal models. These results further support the 3d Mind Model by suggesting that rationality, social impact, and valence are the best set of dimensions for predicting neural representations of mental states out a broad array of alternatives.

Next, we examined the performance of combinations of dimensions (i.e. “models”). The 3d Mind Model again emerged as the best performing set. This model, corresponding to PCs 1–3, was selected as the optimal model 32.90% of the time. The next best model consisted of PCs 1–3, plus PC 11, and was selected as optimal 10.11% of the time – less than a third as often as the 3d Mind Model. All other combinations of PCs were selected less than 10% of the time. Most of the best performing models had a high degree of overlap with PCs 1–3, as reflected in the individual dimension results in the previous paragraph. In other words, none of the 58 dimensions (or 11 PCs) we considered could reliably add predictive power to the 3d Mind Model. Thus, these results indicate that rationality, social impact, and valence remain the best explanation for the neural representation of mental states, even when many other candidate dimensions are considered.

Finally, we evaluated the performance of the optimal model. The model selection procedure compared the 3d Mind Model against other potential combinations of dimensions and found that the former provided the best explanation of the neural representation of mental states. However, it is always possible that the 58 candidate dimensions were not exhaustive. That is, we might have missed some important dimension, despite considering a broad set of possible candidates. To assess how comprehensively the 3d Mind Model captures mental state representations, we conducted a noise ceiling analysis, comparing it to the maximal performance possible given the reliability of data. Results indicated that rationality, social impact, and valence together explain 80.75% of the reliable variance in neural pattern similarity. This high level of explained variance indicates that the 3d Mind Model is a highly explanatory account of how the brain represents mental states.

3.3 Discussion

In Study 2, we examined whether the 3d Mind Model is the optimal model for explaining mental state representation. A principal component analysis (3.1) indicated that 11 orthogonal components could explain 58 different candidate mental state dimensions. The first 3 of these 11 mapped onto valence, rationality, and social impact respectively, indicating that these dimensions explain much of the variance across a wide array of psychological features across mental states. Next, a cross-validated model selection procedure (3.2) selected the best combination of the 11 components for predicting neural representations of mental states. Valence, rationality, and social impact were included in the vast majority of optimally performing models across multiple iterations of this procedure. Moreover, these three dimensions together were the most frequently selected optimal model by a wide margin, indicating that the 3d Mind Model provides a better explanation of mental states than any other set of dimensions considered. Finally, we evaluated this optimal model

using a noise ceiling analysis, which indicated that the 3d Mind Model could explain over 80% of the reliable variance in neural pattern similarity. Together these results indicate the 3d Mind Model outperforms a broad array of alternatives and is a highly explanatory model of mental state representation.

4. Study 3: Generalization to behavior and text semantics

The 3d Mind Model has been evaluated primarily in terms of its ability to explain neural representations of mental states. Although this measure has many merits, it is important to demonstrate that the 3d Mind Model generalizes to other measures of how people think about thoughts and feelings. To complement the neural findings in Studies 1 and 2, Study 3 examined the ability of this model to explain other measures of how people think about mental states. We focused on two such measures: people's explicit judgements about mental states, and the semantics of mental state word use. In the former case, we examined a mental state similarity judgement task, in which participants repeatedly indicated which of two mental states was more similar to a third. This task provided an explicit measure of the conceptual similarity between mental states, which we attempted to predict using the 3d Mind Model. Second, in the text analysis, we relied on an existing measure of semantic similarity, in the form of a pre-trained word vector embedding (Bojanowski, Grave, Joulin, & Mikolov, 2017). This embedding consists of numerical vectors which have been trained to represent word semantics. If the 3d Mind Model generalizes to these alternative measures of mental state representation, this result would support the broad applicability of this model.

4.1 Study 3A: Behavioral measure of mental state representation

4.1.1 Materials and Methods

4.1.1.1 Participants: Participants (N = 124) were recruited to participate in a mental state similarity judgement task on MySocialBrain.org. Sample size was determined by the volume of spontaneous web traffic to the site. Four participants were excluded due to reporting that they experienced technical issues during the experiment. One participant was excluded for responding in < 500 ms on more than 90% of trials, indicating non-compliance with the task. These exclusions left a final sample size of 119 people (73 female, 32 male, 3 other, 11 declined to state sex; mean age = 31, age range = 18–67). Participants provided informed consent in a manner approved by the Institutional Review Board at Princeton University.

4.1.1.2 Experimental procedure: Participants completed a brief demographic survey, and then engaged in a similarity judgement task. On each trial, participants were presented with three states randomly drawn from a total set of 166. One of these states was presented as a reference, the other two as choices (referred to hereafter as "left" and "right"). Participants were instructed to indicate (via button press or mouse click) which of the two choice mental states they thought was more similar to the reference state. For example, on a given trial a participant might be asked to judge whether lethargy is more similar to sadness or anger. After completing 50 trials of this task, participants completed a brief exit survey indicating their enjoyment of the task and whether they had experienced technical issues or previously completed the same experiment.

4.1.1.3 Behavioral data analysis: Just as we earlier (2.1) tested the 3d Mind Model by measuring its ability to explain neural pattern similarity, here we tested the model's ability to explain human judgements of state similarity. Specifically, we predicted that mental states that are closer to each other on the dimensions of rationality, social impact, and valence would be judged as more similar by participants. We began by excluding trials (77 out of 5950) with implausibly short response times (<500 ms). We then analyzed participants' remaining responses using a generalized (binomial) mixed effects model. This model regressed participants' choices (left state versus right state) onto predictors generated from each of the three dimensions. We generated these predictors by taking the absolute difference between the reference and the right choice on a given dimension, and then subtracting away the absolute difference between the reference and the left choice. In other words, these regressors encoded whether the left or the right choice was closer to the reference state along a given dimension. Note that in this analysis and the subsequent text analysis (4.2) we used the scores from ratings in Study 1, not the scores from the new PCA conducted in this investigation (3.1). In addition to the fixed effects, we included random effects to control for dependencies in the repeated measures design. Specifically, we included random intercepts for participant, reference state, left choice state, and right choice state. We also attempted to include random slopes within participant for the three fixed effects. However, we were forced to drop these terms to allow the model to converge. This mixed effects model allowed us to conduct null hypothesis significance testing on the effects of each of three dimensions on participants' choices.

In addition, we estimated the out-of-sample predictive performance of the 3d Mind Model as a whole (i.e., combining all three dimensions). To do so, we used a leave-one-participant-out cross-validation procedure. In this procedure, a binomial regression was trained on the data from all but one participant. The model consisted of participants' choices as the dependent variable, and the same independent variables as the fixed effects in the mixed effects model (i.e., the 3d Mind Model predictors). No random effects were included. The fitted model was then used to predict the choices made by the one left-out participant. These predictions were compared to the actual responses given by that participant to compute the accuracy of the model. This process was then repeated, leaving out each participant in turn. The 166 states presented in the task implied 2M possible unique combinations, of which we sampled less than 1% across all participants. As a result, cross-validating by participant also effectively cross-validated with respect to combinations of mental states. We computed the mean accuracy across participants to estimate the overall performance of the model.

4.1.2 Results—Rationality, social impact, and valence were each significant predictors of mental state similarity judgements. Valence had the largest effect: for every standard deviation of difference between the two choice states on valence, the odds of choosing the more proximal state increased 2.4-times ($z = 24.69$, $p = 3.9 \times 10^{-134}$). Social impact had the second largest effect (odds ratio = 1.43, $z = 11.28$, $p = 1.5 \times 10^{-29}$) followed by rationality (odds ratio = 1.19, $z = 5.50$, $p = 3.9 \times 10^{-8}$). Together, the three dimensions of the 3d Mind Model could predict 65.7% of participants' choices out-of-sample (chance = 50.8%).

4.2 Study 3B: Linguistic measure of mental state representation

4.2.1 Materials and Methods—The semantics of word use in natural language offer another important criterion against which to validate the 3d Mind Model. Using word vector embeddings – a recent advance in computational text analysis – researchers can now extract the semantics of words from the statistical properties of large corpora of text (Mikolov, Chen, Corrado, & Dean, 2013). The semantics of each word are represented as a high dimensional vector – in present case, a set of coordinates in a 300d space. The locations of words in this space are determined by training a model to predict words in text based on other nearby words. The closer two words are to each other within this embedding space, the more similar their meaning. By applying this measure specifically to mental state words, we generated a convergent measure of how people think about mental states. This measure has three important advantages. First, it is naturalistic, in the sense that it is derived from naturally occurring text, rather than text or other behavior elicited in an experiment. Second, it is implicit (Caliskan, Bryson, & Narayanan, 2017), and thus less susceptible to demand characteristics. Third, it provides a broader window into human nature because – in the case of the particular embedding that we use – it draws upon most of the English language text on the internet, rather than the responses of a relatively small number of participants. For these reasons, validating the 3d Mind Model against word vector semantics provides an excellent way to test its generalizability.

We conducted a text analysis to examine whether rationality, social impact, and valence could explain the semantics of mental state words in natural language. First, we computed the semantic similarity between words using fastText, a recent word vector embedding library (Bojanowski et al., 2017). We used a set of 300d vectors which had been pre-trained on the Common Crawl, a set of 600 billion word tokens comprising of most of the English language text on the internet. From this version of fastText, we extracted 166 word vectors corresponding to the mental states rated in Study 1. We then estimated the semantic similarity between the mental state vectors by computing the Pearson correlations between each pair (Figure S2).

To predict this measure of semantic similarity, we generated regressors based on rationality, social impact, and valence in the same way as we did for the mega-analysis in Study 1 (2.1.4). That is, we computed the absolute difference between each pair of states on each of the dimensions. In this case we then flipped the signs of these values to convert differences to similarities. All three predictors, as well as the semantic similarity, were rescaled to a 0 to 1 range. For descriptive purposes, we computed the zero-order Pearson correlations between each predictor and semantic similarity. We then entered all three of these predictors into a non-negative least squares multiple regression predicting semantic similarity.

Due to the statistical dependencies implied by a correlation matrix (i.e., values in the same row or column are related) we could not rely on the usual parametric test of the regression coefficients. Instead, we performed a version of the Mantel test, modified to accommodate the case of multiple regression rather than correlation. That is, we randomly permuted the rows and columns of the semantic similarity matrix (in same order) while holding the dimension-based predictors constant. We recalculated the coefficients of the regression

10,000 times with different random permutations. The results formed empirical null distributions for each of the three predictors. We then compared the actual coefficients from the unpermuted regression to these null distributions to determine their statistical significance.

4.2.2 Results—Rationality, social impact, and valence were each significant predictors of the semantic similarity between mental state words, as estimated using the fastText word embedding. Rationality had the largest association with semantic similarity ($r = .36$, $b = .23$, $p < .0001$), followed by valence ($r = .21$, $b = .12$, $p < .0001$), and social impact ($r = .12$, $b = .09$, $p < .0001$). Biplots in Figure S3 provide an indication of how these psychological dimensions map onto vector represents of mental state words. These results indicate that the same three dimensions which predict the neural and behavioral measures of mental states similarity also predict how mental state words are used in natural language.

The fastText embedding used in Study 3B also allowed us to test whether the 3d Mind Model is specific to the social domain (see Supplementary Information). This embedding places mental state words in a common 300d space along with all other words. We trained a model to predict the 3d Mind Model dimensions from the fastText dimensions for a subset of mental states, and then predict scores for a separate subset of mental states as well as non-mental state terms (objects). We found rationality, social impact, and valence expressed much greater variance within mental states than objects. This suggests that these dimensions preferentially span the social domain, despite any apparent resemblance to dimensions that capture domain-general semantics (Osgood et al., 1957).

4.3 Discussion

The results of Study 3 demonstrate that the 3d Mind Model generalizes beyond neural measures of mental state representation to explain behavioral and semantic measures of how people think about mental states. In Study 3A, participants made explicit judgements about the similarity between mental states. Rationality, social impact, and valence each contributed to predicting these judgements. In Study 3B, the semantics of mental state words were estimated using fastText, a word vector embedding. Again, all three dimensions of the 3d Mind Model significantly predicted this measure of mental state representation. Together, these results provide evidence for the broad applicability of the 3d Mind Model.

5. General discussion

How do people represent the thoughts and feelings of other people? The results presented here provide strong evidence that people do so by arraying these mental states on three psychological dimensions: rationality, social impact, and valence. This 3d Mind Model was validated via a mega-analysis of four fMRI studies (2.), as well as convergent behavioral (4.1) and linguistic (4.2) measures of mental state representation. All three dimensions were significant, unique predictors of mental state representations. The 3d Mind Model offered the optimal combination of a broad array of candidate dimensions, explaining over 80% of reliable variance across fMRI studies (Section 3). Together the results of this investigation provide clear and comprehensive support for the 3d Mind Model of mental state representation.

The present results indicate that the three-dimensional model is a highly explanatory characterization of the shared concepts people use to represent mental states. Over 80% of the reliable variance in neural pattern similarity can be explained by just these three dimensions. This makes the 3d Mind Model both outstandingly powerful and parsimonious by the standards of typical social psychological effects. For instance, a recent meta-scientific investigation of effect sizes in social psychology found that correlation coefficients of .10, .25, .40 should be descriptively considered small, medium and large effect sizes, respectively (Lovakov & Agadullina, 2017). Explaining 80% of variance would translate to a correlation coefficient of ~.9. Thus, the 3d Mind Model of more than doubles the size of large effects in its literature, suggesting that it provides a compelling explanation for how the brain represents mental states.

In addition to its explanatory power, the 3d Mind Model is also broadly applicable. It explains neural representations of mental states, behavioral judgements of mental states, and the semantics of mental state word use in natural language. This generalizability suggests that rationality, social impact, and valence influence both explicit judgements and implicit cognition. Word embeddings like the fastText vectors we used here encode the semantics by predicting words based on their context (Bojanowski et al., 2017). The ability of the 3d Mind Model to predict these semantics suggests that the 3d Mind Model could also explain how people choose which mental state words to use in written or spoken communication. Together, this convergent evidence from multiple measurement modalities provides stronger support for the 3d Mind Model than any one modality could on its own.

The 3d Mind Model is also robust to a wide range of low-level design features of the fMRI studies. The four studies we consider here varied in terms of task timing, response type, the particular mental states presented, and the scenarios paired with them. These studies also varied on a number of incidental features, such as imaging parameters, participants, and scanner hardware. Although these design features account for some of the heterogeneity we observe in effect sizes across studies, analysis of variance components suggests that the study-specific effects are relatively small (see Supplementary Information). Moreover, we still observe robust effect when mega-analyzing these studies together (Figure 1B). Although the fMRI studies presented here do have many features in common, the robustness of results across their variations does suggest that low-level task features are unlikely to account for the explanatory power of the 3d Mind Model.

The 3d Mind Model generalizes across many different tests of mental state representation, but not to other categories of stimuli. The dimensions of rationality, social impact, and valence bear a resemblance to the more domain-general dimensions of potency, activity, and evaluation, respectively, from Osgood’s semantic differential (Osgood et al., 1957). This parallel raises the question of whether the 3d Mind Model is specific to minds, or instead reflects a more general characterization of semantics. For example, to what extent is social valence, as we consider here, the same construct as the general semantic valence from Osgood’s theory, among others? To address this question, we used text analyses to test whether the 3d Mind Model applies to the domain of physical objects to the same extent that it spans the domain of mental objects (see Supplementary Information). We found strong

evidence that rationality, social impact, and valence preferentially span mental state space. This result suggests the 3d Mind Model is indeed specific to the social domain.

The 3d Mind Model explains much of how people think about mental states. However, there are several limitations to its explanatory power worth noting (Saxe, 2018). First, the 3d Mind Model is a model of shared concepts of mental states. That is, it can explain what is common across studies and brains, but not what varies. The model cannot currently make different predictions about how different people represent mental states differently from one another. For example, one person's concept of happiness might differ from another's in subtle yet important ways. The 3d Mind Model cannot predict when, or between whom, such differences will arise. It is a model of people's shared understanding of mental states, rather than of each of their idiosyncratic understandings. Thus, when we report that this model explains 80% of the variance in mental state representation, it should be understood that it explains 80% of the shared variance, not the idiosyncratic variance. More research is needed on individual differences in mental state representation in order to examine the adequacy of this model in that regard. Cross-cultural extension of this model should also be a high priority for future studies, as the present investigation cannot speak to this form of generalization either (Saxe, 2018).

Second, the 3d Mind Model offers a framework for mental state concepts – how people think about mental states – and not actual emotion experiences, as felt by people in the moment. Everyday personal experiences of emotion, as well as "live" perceptions of others' emotions, differ from the emotion concepts described by the model. Experienced emotions are appraised in the light of contextual factors (Barrett, Mesquita, & Gendron, 2011), factors which did not vary systematically in the studies presented here. A recent fMRI investigation (Skerry & Saxe, 2015) found that contextual appraisal dimensions significantly outperformed the core affect dimensions of valence and arousal (Russell, 1980, 2003) in explaining patterns of brain activity elicited by reading emotionally evocative scenarios. The mismatch between people's conception of mental states and emotions embedded in real life may help explain why we find that the 3d Mind Model performs so well at describing concepts, but 27 dimensions appear optimal for describing experience (Cowen & Keltner, 2017).

Third, the 3d Mind Model cannot explain how people account for the propositional content of mental states (Saxe, 2018). For example, the model can explain the abstract representation of the state of desire, aggregated over many instances. However, this state inherently implies a propositional object: what do people desire? There is a very meaningful difference between "desire for a tenure-track job" and "desire for a piece of chocolate," but the 3d Mind Model cannot yet account for this distinction. It is possible that such distinctions may in some cases reduce to small perturbations around the average coordinates of a state in the 3d space (e.g., desiring a job may drag the representation towards the more rationality pole of the space). However, in cases where the meaning of a state depended entirely on its propositional content, the perturbations could be so large as the render the center meaningless. That said, of the states we investigate here, even those which seem to be highly dependent on propositional content – like belief and desire – still possess a conceptual core that is shared across propositions. This is reflected by the fact that these states elicit reliable

patterns of brain activity, consistent behavioral judgements, and meaningful semantics in text, much like their less propositional cousins. Although the 3d Mind Model may or may not describe the propositional variations within each state, it is nonetheless useful in describing the core concepts which states share across propositions. States with propositional content might also vary along additional, as-yet unidentified dimensions not included in the model. Other models of theory of mind, such as inverse planning models (Baker, Saxe, & Tenenbaum, 2009), may deal more easily with propositional content. However, it can be difficult to scale such models beyond relatively simplistic or tightly controlled paradigms. Determining how to hybridize such models with the 3d Mind Model might help address the limitations of each. Thus, while the present results provide strong support for the 3d Mind Model, this model still has a number of clear limitations (Saxe, 2018). Further research and model development should be conducted to address these limitations.

Why should social perceivers represent mental states with the 3d Mind Model? The model gives perceivers a way of describing and comparing others' thoughts and feelings, but how can this functionality be applied to practical social problems? In earlier work, we speculated that these dimensions might form the basis of a social evaluation system which could help identify allies or threats (Tamir et al., 2016). Specifically, social impact could encode how likely someone is to engage with you in a socially meaningful way; valence could encode whether that engagement would likely be positive or negative; and rationality could encode the competence with which the other person engages. Support for this functional interpretation of mental state representation already exists in the literature. Research on the "face in the crowd" effect suggests that there is an attentional advantage to perceiving angry faces (Hansen & Hansen, 1988; Öhman, Lundqvist, & Esteves, 2001; Pinkham, Griffin, Baron, Sasson, & Gur, 2010). This is highly consistent with the functional account of the 3d Mind Model, since anger is a state high on social impact and negative on valence, and thus – by our account – threatening. In another set of studies, researchers found that mentalizing about certain types of states might differentially promote certain social behaviors. Specifically, explicitly mentalizing about states in the positive and emotional regions of the 3d Mind causally increases feelings of social connection, and social behaviors such as information sharing (Baek, Tamir, & Falk, 2019). More work needs to be done to fully characterize the social function of each individual dimension, or any combinations thereof.

We have since identified another potential use for these dimensions: mental state prediction. In previous research, we investigated how people accurately predict the transitions between mental states (Thornton & Tamir, 2017). We used first measured how people actually transition from state to state using experience-sampling data – in which people periodically reported on their mental states as they went about their daily lives. There is a lot of agreement across people about how emotions unfold over time: some transitions are reliably rated as highly likely, and others not as likely. For example, people who report feeling joyful at one moment in time are much more likely to later report feeling gratitude than are people who initially report feeling disgust. We can find meaningful structure in these transitions. Specifically, the closer two states are in the 3d Mind space, the higher their transition likelihood. Proximity (or distance) on these dimensions reflects *actual* emotion dynamics. Importantly, people also use these dimensions to make predictions about others' emotion

transitions. The closer two states are to each other in 3d Mind space, the more people predict that others will experience a transition between them. Thus, by using these dimensions, people are able to make accurate predictions about others' emotions. This suggests that the very way in which we think of mental state may be organized around the goal of prediction. We have since proposed that this may be a general feature of social prediction: representations of trait, states, actions, and situations may all be organized along dimensions that aid prediction (Tamir & Thornton, 2018).

In the current study, we use neural data to show that the brain uses the 3d Mind Model to represent mental states, such that the closer two states are in the 3d Mind space, the more similar their neural patterns. This neural similarity also supports mental state prediction (Thornton, Weaverdyck, & Tamir, 2019). We have previously shown that the social brain network – and medial prefrontal and parietal cortices in particular – encodes transition likelihood. In fact, the 3d Mind Model predicts neural pattern similarity within the same regions that encode state transition probabilities. Indeed, the 3d Mind dimensions partially explain the relation between transition probability ratings and neural pattern dissimilarity. This suggests that the brain may use the 3d Mind Model to predict the social future. The 3d Mind Model may thus support a form of social predictive coding (Friston & Kiebel, 2009; Koster-Hale & Saxe, 2013).

Another question the present data allow us to consider is whether rationality, social impact, and valence are appropriately named. In the present study, participants read definitions based on these names and rated mental states accordingly. These ratings were among the 58 subjected to a PCA in Study 2. By examining the loadings of these ratings onto the extracted components, we could assess how well these names match up with the underlying meaning of each dimension. The results appear promising for the rationality and valence dimensions. Explicit ratings of these two dimensions had the highest loadings out of all 58 dimensions onto PCs 2 and 1 respectively (Table S1). This suggests that these names may indeed be the best descriptors for these dimensions, at least out of those we have examined. Moreover, these new PCs were highly correlated with the PCs which we called rationality and valence in our earlier study (Table 3). The case is somewhat less clear-cut when it comes to social impact. Human ratings of social impact loaded highly onto PC 3 ($r = .72$). However, several other rated dimensions had even larger loadings on PC 3: high arousal ($r = .92$), low arousal ($r = -.92$), beginning abruptness ($r = .84$), and ending abruptness ($r = .82$). This suggests that reverting to the use of the term arousal – originally from the circumplex model (Russell, 1980) – may ultimately prove more intuitive for raters of this dimension. That said, PC 3 was highly correlated with the dimension we call "social impact" (r = .90) in previous research (Tamir et al., 2016). This suggests that the meaning of this dimension remains essentially the same, despite the question of how best to name it.

The present investigation aimed to provide a comprehensive answer to the question of how people think about thoughts and feelings. The results strongly support the 3d Mind Model of mental state representation, consisting of rationality, social impact, and valence. This model is a robust, and highly explanatory explanation of neural representations of mental states, outperforming all of a broad array of alternative models. It also generalizes from fMRI to behavior and text semantics. On this basis, we conclude that the 3d Mind Model can provide

a fruitful basis for future research into how people understand each other's minds. Such future research could help to reveal the nature of functional impairments in theory of mind, such as in autism (Baron-Cohen, 1997; Frith & Happé, 1994). The 3d Mind Model could also inform research in affective computing by improving (or at least, making more human) the logic by which artificially intelligent systems judge and react to the states of human agents (Picard & Klein, 2002). Such possibilities highlight the promising applications which may follow from a solid understand of how the brains represents minds.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Jordan Rubin-McGregor for her assistance. This work was supported by NIMH grant R01MH114904 to D.I.T.

7. References

Ashburner J (2007). A fast diffeomorphic image registration algorithm. Neuroimage, 38(1), 95–113. [PubMed: 17761438]

Baek E, Tamir D, & Falk E (2019). Considering Others' Mental States Causally Increases Feelings of Social Bonding and Information Sharing.

Baker CL, Saxe R, & Tenenbaum JB (2009). Action understanding as inverse planning. Cognition, 113(3), 329–349. [PubMed: 19729154]

Baron-Cohen S (1997). Mindblindness: An essay on autism and theory of mind. MIT press.

Barrett LF, Mesquita B, & Gendron M (2011). Context in emotion perception. Current Directions in Psychological Science, 20(5), 286–290.

Bojanowski P, Grave E, Joulin A, & Mikolov T (2017). Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135–146.

Britton JC, Phan KL, Taylor SF, Welsh RC, Berridge KC, & Liberzon I (2006). Neural correlates of social and nonsocial emotions: An fMRI study. NeuroImage, 31(1), 397–409. [PubMed: 16414281]

Caliskan A, Bryson JJ, & Narayanan A (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183–186. [PubMed: 28408601]

Cowen AS, & Keltner D (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proceedings of the National Academy of Sciences, 114(38), E7900–E7909.

Cuddy AJ, Fiske ST, & Glick P (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. Advances in Experimental Social Psychology, 40, 61–149.

De La Vega A, Chang LJ, Banich MT, Wager TD, & Yarkoni T (2016). Large-scale meta-analysis of human medial frontal cortex reveals tripartite functional organization. Journal of Neuroscience, 36(24), 6553–6562. [PubMed: 27307242]

Eldar E, Rutledge RB, Dolan RJ, & Niv Y (2016). Mood as representation of momentum. Trends in Cognitive Sciences, 20(1), 15–24. [PubMed: 26545853]

Fiske S, Cuddy A, Glick P, & Xu J (2002). A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. Journal of Personality and Social Psychology, 82(6), 878–902. [PubMed: 12051578]

Frijda NH (2004). Emotions and action. Feelings and Emotions: The Amsterdam Symposium, 158–173. Cambridge, UK: Cambridge University Press.

Friston K, & Kiebel S (2009). Predictive coding under the free-energy principle. Philosophical Transactions of the Royal Society B: Biological Sciences, 364(1521), 1211–1221.

Frith U, & Happé F (1994). Autism: Beyond “theory of mind.” Cognition, 50(1–3), 115–132. [PubMed: 8039356]

Hackforth R (1952). Plato: Phaedrus (Vol. 119). Cambridge University Press.

Hansen CH, & Hansen RD (1988). Finding the face in the crowd: an anger superiority effect. Journal of Personality and Social Psychology, 54(6), 917–924. [PubMed: 3397866]

Heckhausen H, & Gollwitzer PM (1987). Thought contents and cognitive functioning in motivational versus volitional states of mind. Motivation and Emotion, 11(2), 101–120.

Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, & Smith SM (2012). Fsl. Neuroimage, 62(2), 782–790. [PubMed: 21979382]

Kahneman D (2003). Maps of bounded rationality: Psychology for behavioral economics. American Economic Review, 1449–1475.

Khaligh-Razavi S-M, Henriksson L, Kay K, & Kriegeskorte N (2017). Fixed versus mixed RSA: Explaining visual representations by fixed and mixed feature sets from shallow and deep computational models. Journal of Mathematical Psychology, 76, 184–197. [PubMed: 28298702]

Khaligh-Razavi S-M, & Kriegeskorte N (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Computational Biology, 10(11).

Konkle T, & Oliva A (2012). A real-world size organization of object responses in occipitotemporal cortex. Neuron, 74(6), 1114–1124. [PubMed: 22726840]

Koster-Hale J, & Saxe R (2013). Theory of mind: a neural prediction problem. Neuron, 79(5), 836–848. [PubMed: 24012000]

Kriegeskorte N, Mur M, & Bandettini PA (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience, 2.

Lovakov A, & Agadullina ER (2017). Empirically Derived Guidelines for Interpreting Effect Size in Social Psychology. PsyArXiv.

Mikolov T, Chen K, Corrado G, & Dean J (2013). Efficient estimation of word representations in vector space. ArXiv Preprint ArXiv:1301.3781.

Nichols TE, & Holmes AP (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. Human Brain Mapping, 15(1), 1–25. 10.1002/hbm.1058 [PubMed: 11747097]

Öhman A, Lundqvist D, & Esteves F (2001). The face in the crowd revisited: A threat advantage with schematic stimuli. Journal of Personality and Social Psychology, 80, 381–396. [PubMed: 11300573]

Osgood CE, Suci GJ, & Tannenbaum PH (1957). The measurement of meaning. Oxford, England: University of Illinois Press.

Owen AB, & Perry PO (2009). Bi-cross-validation of the SVD and the nonnegative matrix factorization. The Annals of Applied Statistics, 3(2), 564–594.

Pessoa L (2008). On the relationship between emotion and cognition. Nature Reviews Neuroscience, 9(2), 148. [PubMed: 18209732]

Picard RW, & Klein J (2002). Computers that recognise and respond to user emotion: theoretical and practical implications. Interacting with Computers, 14(2), 141–169.

Pinkham AE, Griffin M, Baron R, Sasson NJ, & Gur RC (2010). The face in the crowd effect: anger superiority when using real faces and multiple identities. Emotion (Washington, DC), 10(1), 141–146.

Posner J, Russell JA, & Peterson BS (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. Development and Psychopathology, 17(03), 715–734. [PubMed: 16262989]

Russell JA (1980). A circumplex model of affect. Journal of Personality and Social Psychology, 39(6), 1161–1178.

Russell JA (2003). Core affect and the psychological construction of emotion. Psychological Review, 110(1), 145–145. [PubMed: 12529060]

Russell JA, & Mehrabian A (1977). Evidence for a three-factor theory of emotions. Journal of Research in Personality, 11(3), 273–294.

Saxe R (2018). Seeing Other Minds in 3D. Trends in Cognitive Sciences, 22(3), 193–195. [PubMed: 29482823]

Skerry AE, & Saxe R (2015). Neural representations of emotion are organized around abstract event features. Current Biology, 25(15), 1945–1954. [PubMed: 26212878]

Smith CA, & Ellsworth PC (1985). Patterns of cognitive appraisal in emotion. Journal of Personality and Social Psychology, 48(4), 813–838. [PubMed: 3886875]

Sung YJ, Schwander K, Arnett DK, Kardia SL, Rankinen T, Bouchard C, … Rao DC (2014). An empirical comparison of meta-analysis and mega-analysis of individual participant data for identifying gene-environment interactions. Genetic Epidemiology, 38(4), 369–378. [PubMed: 24719363]

Tamir DI, & Thornton MA (2018). Modeling the predictive social mind. Trends in Cognitive Sciences, 22(3), 201–212. [PubMed: 29361382]

Tamir DI, Thornton MA, Contreras JM, & Mitchell JP (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. Proceedings of the National Academy of Sciences, 113(1), 194–199.

Thornton MA, & Tamir DI (2017). Mental models accurately predict emotion transitions. Proceedings of the National Academy of Sciences, 114(23), 5982–5987.

Thornton MA, Weaverdyck ME, Mildner JN, & Tamir DI (2019). People represent their own mental states more distinctly than those of others. Nature Communications, 10(2117). 10.1038/s41467-019-10083-6

Thornton MA, Weaverdyck ME, & Tamir DI (2019). The social brain automatically predicts others' future mental states. Journal of Neuroscience, 39(1), 140–148. [PubMed: 30389840]

Winkler AM, Ridgway GR, Douaud G, Nichols TE, & Smith SM (2016). Faster permutation inference in brain imaging. NeuroImage, 141, 502–516. [PubMed: 27288322]

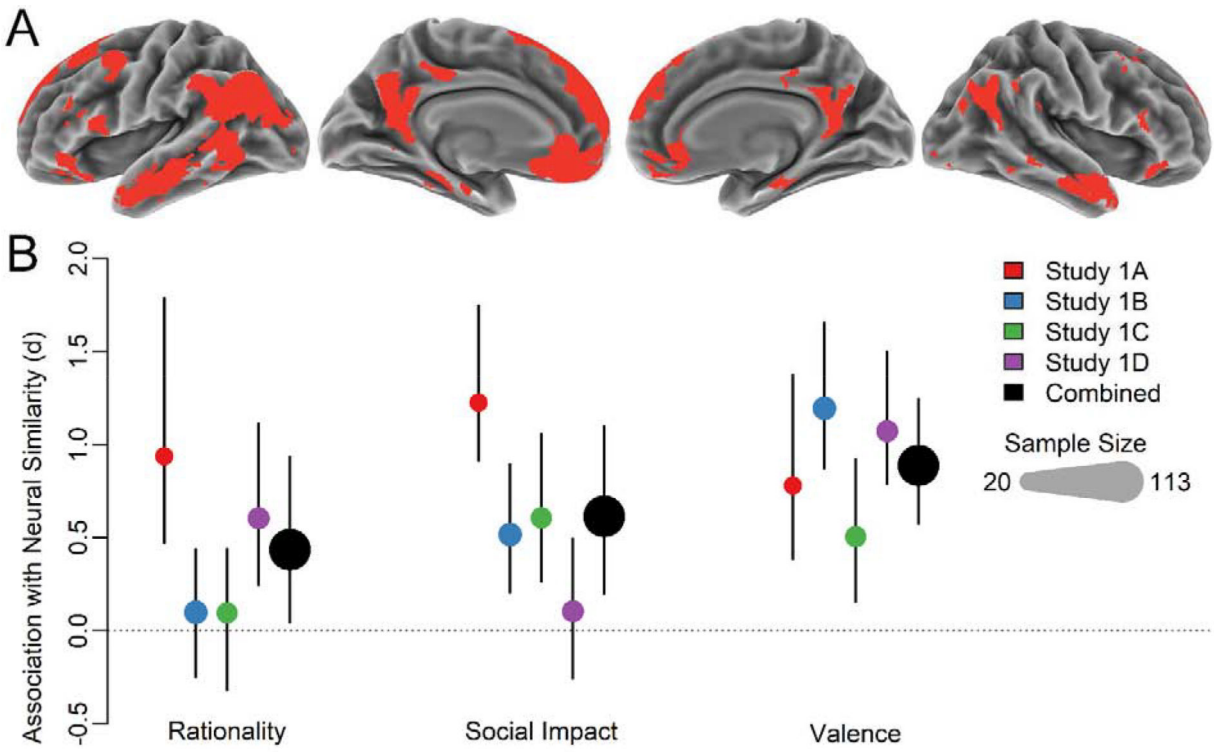

Figure 1. Mega-analysis of the dimensions of mental-state representation.

A) The social brain network. The orange regions denote the 10,216 voxels included in the network-based fMRI analyses. These voxels were selected to be sensitive to mental state representation using an independent dataset. B) The dimensions of rationality, social impact, and valence explain patterns of neural activity within the social brain when thinking about mental states. Circles indicate the effect size (Cohen's *d*) of the association between each 3d Mind dimension and neural pattern similarity across Studies 1A-D. Circle area reflects sample size. Error bars indicate 95% CIs.

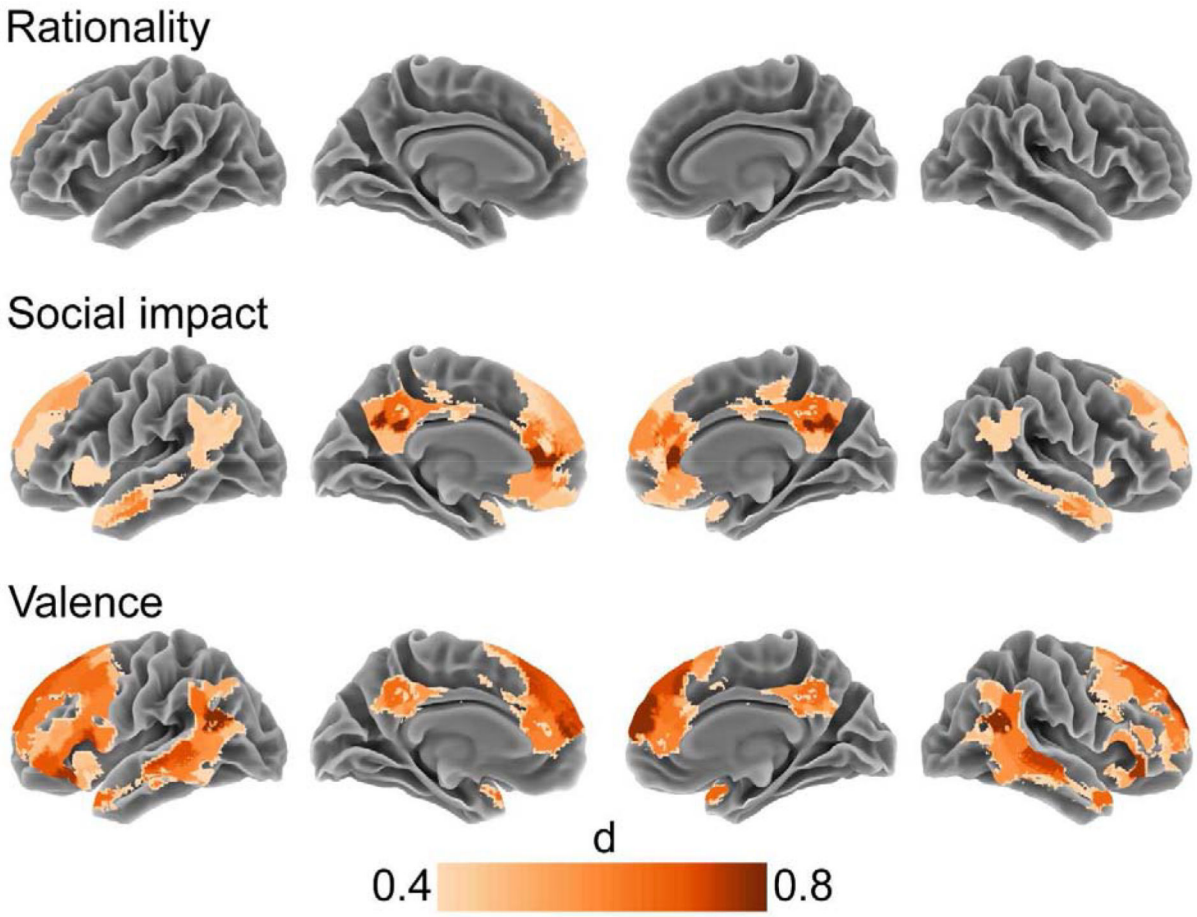

Figure 2. Parcel-based mega-analysis results.

Wholebrain analysis, across 200 parcels, reveal regions where the 3d Mind Model dimensions explain patterns of neural activity when thinking about mental states, across all participants in Studies 1A-D.

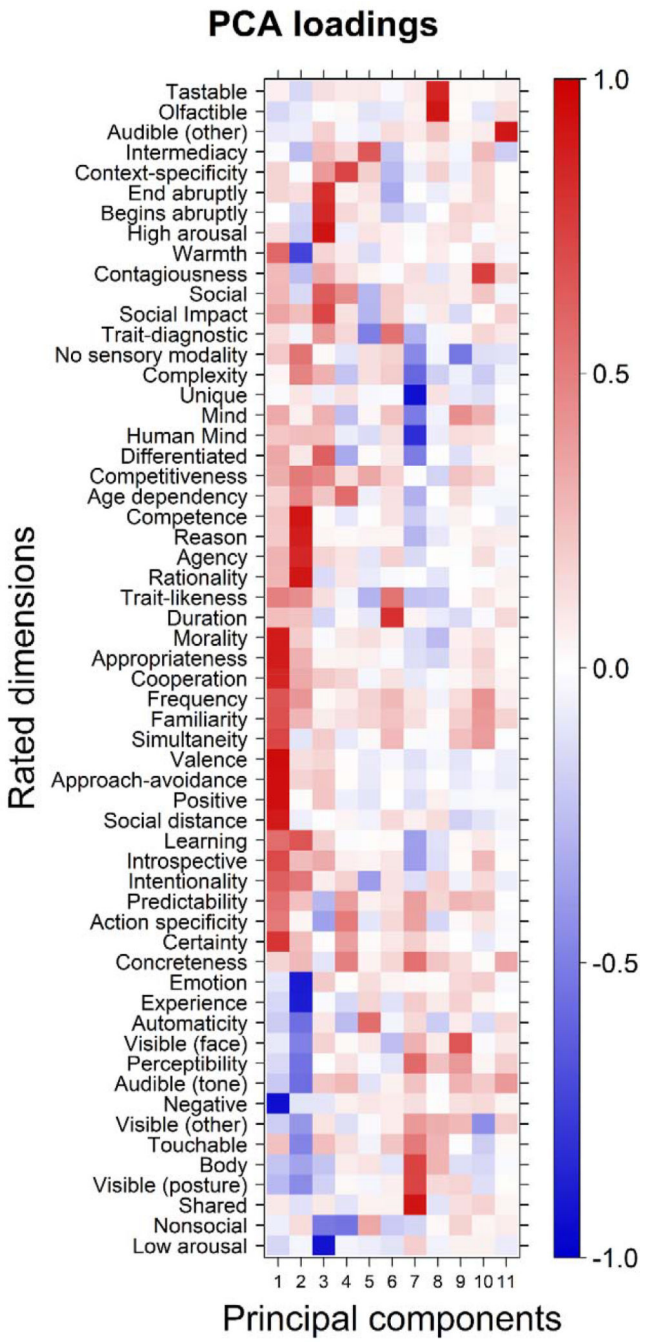

Figure 3. PCA loadings.

The heatmap represents the loadings after varimax rotation of an 11-component PCA conducted across all 58 rated dimensions of mental state representation. These dimensions (y-axis) are grouped based on their loadings, with dimensions with similar loadings shown next to each other. The loadings can be interpreted as the correlation between each component (x-axis) and rated dimension. The first three components map closely onto valence, rationality, and social impact, respectively.

Table 1.

Design characteristics of fMRI studies

Study	N	States	Scenarios	Task	Targets	Runs	Trials/Run	Trial length
Study 1A	20	60	Text	Choice	Generic other	16	60	4.75 s
Study 1B	35	25	Text	Choice	Self, close, far	12	75	4.5 s
Study 1C	28	15	Text	Rating	Generic other	4	225	2.5 s
Study 1D	30	30	Images	Choice	Self, far	12	60	4.2 s

Note: Scenarios presented as text consisted of short (~3–5 word) statements such as “seeing the pyramids” or “watching a meteor shower” that corresponded to a given mental state (e.g., awe); Images consisted of photos and drawings that conveyed similar content. The choice task consisted of selecting which of two scenarios would elicit more of a given state in a target person. The rating task consisted of rating the extent to which one such scenario would elicit a given state in a target person on a Likert-type scale. All trials were interspersed with fixation periods, including random jitter equal to 1/4th the total length of experiment, except in Study 3, which followed a continuous carryover design.

Table 2.

Imaging procedure of fMRI studies.

Study	Scanner	Head coil	Resolution (mm)	TR (ms)	TE (ms)	Flip angle (°)
Study 1A	Trio	32 channel	2.5 × 2.5 × 2.5	2500	30	90
Study 1B	Prisma	64 channel	2.0 × 2.0 × 2.0	2250	30	70
Study 1C	Skyra	64 channel	2.0 × 2.0 × 2.0	1500	32	70
Study 1D	Prisma	64 channel	2.0 × 2.0 × 2.0	1400	32	70

Note: All functional scans consisted of echo-planar images (EPIs) with wholebrain coverage.

Table 3.

Correlations between PCs and Rationality, Social Impact, and Valence

PC	Rationality	Social Impact	Valence
1	.13	.04	.96
2	.97	−.07	−.12
3	.02	.90	.00
4	.06	.28	−.14
5	−.08	−.11	−.08
6	.08	.21	−.02
7	−.02	.06	−.05
8	−.07	.03	.01
9	−.07	.04	.02
10	−.02	.09	.04
11	.00	.01	−.04

Note: Rationality, Social Impact, and Valence in this table represent principal components from earlier research (Tamir et al., 2016), not the human-rated versions shown in Table S1.