Published in final edited form as:

Neuroimage. 2021 September ; 238: 118258. doi:10.1016/j.neuroimage.2021.118258.

The representational structure of mental states generalizes across target people and stimulus modalities

Miriam E. Weaverdyck[a,1,*], **Mark A. Thornton**[a,2], **Diana I. Tamir**[a,b]

[a]Department of Psychology, Princeton University, Princeton, NJ 08544, United States

[b]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544, United States

Abstract

Each individual experiences mental states in their own idiosyncratic way, yet perceivers can accurately understand a huge variety of states across unique individuals. How do they accomplish this feat? Do people think about their own anger in the same ways as another person's anger? Is reading about someone's anxiety the same as seeing it? Here, we test the hypothesis that a common conceptual core unites mental state representations across contexts. Across three studies, participants judged the mental states of multiple targets, including a generic other, the self, a socially close other, and a socially distant other. Participants viewed mental state stimuli in multiple modalities, including written scenarios and images. Using representational similarity analysis, we found that brain regions associated with social cognition expressed stable neural representations of mental states across both targets and modalities. Together, these results suggest that people use stable models of mental states across different people and contexts.

Keywords

Social cognition; Mentalizing; fMRI; Representational similarity analysis

1. Introduction

Every person – from a romantic partner to a complete stranger – has a unique mind with unique experiences, expressions, and behaviors. This poses a challenge: How do we understand any one person's idiosyncratic mind? One possibility is that people rely on core

*Corresponding author. mweaverdyck@ucla.edu (M.E. Weaverdyck).

[1]Permanent Address: Department of Psychology, University of California, Los Angeles, 1285 Franz Hall, Los Angeles, CA 90095, United States.

[2]Permanent Address: Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03755, United States.

Author Contributions

M.A.T. and D.I.T. conceptualized the research. M.E.W. conducted the investigation. M.A.T. and M.E.W. contributed software and methodology and conducted the formal analyses. M.E.W. visualized the results. M.E.W. prepared the original draft. All authors contributed to writing the manuscript. D.I.T. supervised the research.

Open Practices Statement

All de-identified data and analysis scripts have been made publicly available via the Open Science Framework and are accessible at https://osf.io/z3xs9/.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.118258.

mental state concepts that are consistent across people. This approach would allow perceivers to generalize their understanding of mental states–including emotions and cognitive states–across people and situations, rather than laboriously constructing unique models for each mind. If people consistently apply the same core concepts when thinking about mental states, we should see this stable core across diverse targets. We should also observe stability regardless of how perceivers receive information about that person (e.g., through pictures or stories). Here, we explore this possibility by examining the stability of mental state representations in the brain across diverse contexts.

Do people approach each person anew, representing their emotional landscape idiosyncratically? This strategy might be appropriate, given that every person is unique. One's best friend has a unique face, traits, behavioral patterns, and relationship history that differentiates them from every other unique individual in this world (e.g., Thornton et al., 2018, Todorov et al., 2007, Trope and Liberman, 2010). Each person's individuality shapes others' understandings of that person's mind and internal states (e.g., Epley, 2008, Harris and Fiske, 2011, Haslam and Loughnan, 2014, Thornton et al., 2019, Thornton et al., 2018). Indeed, we likely consider emotions differently depending on who is experiencing them. A close friend's joy may be seen as much more positive than a stranger's joy (Claypool et al., 2007); an enemy's joy may even be a negative experience (Cikara and Fiske, 2011). These differences in how people consider emotional experiences become particularly stark when comparing one's own experiences to others'. People represent their own experiences more richly than others' (Thornton et al., 2019). These differences in experience may arise, in part, due to the type of information available from each (e.g., interoceptive signals and introspection about one's own mind versus only perceptible cues from others). Thus, when considering someone else's mind, the uniqueness of that person may overwhelm any similarities in how states are experienced in general.

However, an alternative way to make sense of others' unique minds would be to apply core concepts of mental states that generalize across people. That is, people may operate upon concepts of mental states that are largely universal (Thornton et al., 2020). Universality in this case does not mean that everyone has the same understanding of mental state concepts. Rather, it means that people have a similar framework they use to understand mental states; individuals' idiosyncratic representations could still exist within this framework. For example, grief is intense and negative, no matter who is experiencing that grief. If someone tells a friend about a tragic loss, that friend, understanding the nature of grief, can empathize with their experience (de Vignemont and Singer, 2006). Despite any idiosyncrasies that the current situation may impose, both people have a core understanding of what grief *is*. This shared understanding of emotions has been the basis of much affective research, which implies that there are universal emotions that span age, race, and culture (e.g., Barrett et al., 2007, Jackson et al., 2019, Thornton et al., 2020). It may be, then, that everyone uses stable, generalizable concepts of mental states when considering what someone else is experiencing. These concepts would define both the primary features of each state individually, as well as how two states relate to one another (e.g., how similar they are).

Functional magnetic resonance imaging (fMRI) provides a unique method for testing if people use a common mental state model across all targets. Specific regions of the brain,

including the medial prefrontal cortex (MPFC), anterior temporal lobe (ATL), and temporoparietal junction (TPJ), have been shown to encode and process information related to others' minds, including representing others' mental states (Schurz et al., 2020). FMRI allows us to compare neural representations of individual mental state concepts without having to ask the participant, who may not be able to provide an accurate response regarding their understanding of others' minds (Nisbett and Wilson, 1977). By examining how the brain represents mental state concepts across individuals and modalities using these techniques, we can map the abstract representational space comprised of these mental states. These representations can be extracted for each target individual, and then compared across targets to test for commonalities and distinctions in the models.

Even if people use the same model for everyone's mental states, each person offers insight into their state in different ways. For example, we can learn about friends experiencing grief through multiple channels. One friend may text you that they are having a rough day; another friend may look at you with sad, tired eyes; a third friend may cry to you over the phone. Past research using isolated expressions of emotion suggests that the human brain encodes affect similarly across auditory (e.g., voice inflection) and visual (e.g., facial expression, body language) modalities (Bedny et al., 2008, Chikazoe et al., 2014, Peelen et al., 2010). However, to date, much of neuroimaging research on mentalizing has relied primarily on lexical stimuli (e.g., Meyer et al., 2012, Mitchell et al., 2002, Saxe and Kanwisher, 2003, Skerry and Saxe, 2014, Tamir et al., 2016, Theriault et al., 2020, Thornton et al., 2019, Thornton et al., 2019). Since language and social cognition overlap significantly in the brain (and are both largely distinct from perceptual regions), this may contribute to findings that there is some stable structure to neural representations of mental states (Thornton et al., 2019). If humans do use generalizable core concepts of mental states across modalities, then we should see similar neural representations of each mental state across visual and lexical stimuli. That is, watching a friend in agony should generate the same neural representation as reading a description of it. However, since pictures and phrases necessarily contain different types of information, it is also possible that people process these mental states uniquely depending on how the information is presented. Here we test how people represent supramodal mental states to see if core mental state concepts generalize across modalities.

Here, we tested two specific hypotheses about how people represent mental state concepts. First, we expect that the structure of neural representations associated with mental states should remain consistent across multiple targets – from the self to dissimilar others. Second, we expect that neural patterns associated with mental states should remain consistent across modalities – including lexical and pictorial stimuli.

2. Material and methods

To test our hypotheses, we combined data from three fMRI studies in which participants considered different individuals experiencing a variety of mental states using either lexical or pictorial stimuli. By analyzing across studies, we can assess the stability of mental state representations across (a) target people and (b) stimulus modalities. Moreover, this approach

provides a severe test of the generalizability, since participants, fMRI scanners, time, and institutions differ between studies as well.

2.1. Use of published data

In this manuscript, we use datasets that have been analyzed and published in previous manuscripts. Each study's materials and analyses are freely available on the Open Science Framework. Here, we refer to these as Study 1 (https://osf.io/3qn47/; Tamir et al., 2016, Thornton et al., 2018), Study 2 and Study 3 (studies 2 and 1 in https://osf.io/hp5wc/, respectively; Thornton et al., 2019).

2.2. Participants

Participants in Study 1 (N = 20; 16 female, 4 male; 18-27 years, M = 22.7 years) were recruited from the Harvard University Study Pool. The participants in Study 2 (N = 35; 23 female, 12 male; 18-31 years, M = 21 years) and Study 3 (N = 30; 14 female, 15 male, 1 nonbinary; 19-27 years, M = 20 years) were recruited from the Princeton University Credit and Paid Study Pools. The sample size in Study 1 was determined via Monte Carlo simulation parameterized based on effect sizes in previous studies of similar design (Tamir et al., 2016). The sample sizes for Studies 2 and 3 were determined *a priori* to be able to detect the smallest effect found in Study 1 (namely, the relationship between neural representations of mental state and behavioral ratings of valence) with 95% power (Thornton et al., 2019). All participants were right-handed or ambidextrous, fluent English-speakers, reported no history of neurological problems, had normal or corrected-to-normal vision, and were screened for standard MRI exclusion criteria. Participants were compensated with university credit or monetary payment. All data were collected in accordance with the Harvard University Committee on the Use of Human Subjects (Study 1) or Princeton University Institutional Review Board (Studies 2 and 3).

2.3. Experimental design

In all three studies, participants made judgments about a wide range of mental states. On each trial, participants were shown a prompt with a mental state word (e.g., "peacefulness") at the top of the screen, followed by two scenarios pre-tested to elicit that mental state. Participants chose which of the two scenarios, presented in either lexical or pictorial form, would most likely elicit the current mental state in a particular target person.

There were four targets across all three studies. In Study 1, participants chose for a generic other (no target description was given beyond "another person"); in Studies 2 and 3, participants chose for a specific target individual (shown along with the mental state prompt) at various levels of psychological distance from the subject: self, close other, or far other (Fig. 1, Table 1). Self-trials were cued with the word "SELF "; socially close and far targets were cued with the name of that person (e.g., "KATE"). Close targets were nominated by the participant to be likable, familiar, and similar to themselves (i.e., socially close). Far targets were introduced to the participant via a short biographical statement (see Supplement S9 for example bios). The experimenters created these fictional targets to be dissimilar to the participant in terms of their college major, religiosity, political party and attitudes, and

extracurricular activities. Thus, the far target was both unfamiliar and dissimilar to the participant, thereby maximizing the social distance between the target and the participant.

An initial list of 166 mental state words was reduced to an optimized set of 60 states (Tamir et al., 2016). These 60 states were selected for minimal redundancy and to uniformly cover the conceptual space of mental states along many different psychological dimensions (Tamir et al., 2016). All 60 states were used in Study 1; Studies 2 and 3 consisted of subsets of these 60 (Table 1; see Table S1 for full list of state words) that sampled the entire space based on the variance observed in Study 1. Participants saw each target-mental state pair once per run.

Each mental state word and target cue (if applicable) was followed by two scenarios likely to elicit that state in the average person. The two scenarios were drawn from a unique set associated with that mental state. The modality of the scenario stimuli differed by study (Fig. 1): In Studies 1 and 2, scenarios were presented as short text phrases (e.g., “watching the sun rise,” “engaging in meditation”). In Study 3, scenarios were presented as images. These images sometimes consisted of a person experiencing the mental state (e.g., a picture of a person meditating), and sometimes only showed the scenario (e.g., a picture of the sun rising). Participants’ task was to choose which scenario would most likely elicit the mental state in the target person.

There was no “correct” choice, and there were a large number of possible scenario combinations across studies. For this reason, we chose not to analyze scenario choice as related to neural responses in the current project. Response rates were used in the original studies to exclude participants (Tamir et al., 2016, Thornton et al., 2019, Thornton et al., 2018). In the current datasets, response rates were very high (M_{Study1} = 92.44%, SD_{Study1} = 9.19; M_{Study2} = 96.31%, SD_{Study2} = 4.24; M_{Study3} = 96.71%, SD_{Study3} = 4.78) participants chose the right-hand option approximately 50% of the time (M_{Study1} = 51.43%, SD_{Study1} = 5.80; M_{Study2} = 49.75%, SD_{Study2} = 4.57; M_{Study3} = 52.33%, SD_{Study3} = 5.06), and participants had reasonable reaction times (M_{Study1} = 2.35s, SD_{Study1} = 0.35; M_{Study2} = 1.99s, SD_{Study2} = 0.33; M_{Study3} = 1.77s, SD_{Study3} = 0.25), suggesting that participants were alert and focused on the task. We also examined how frequently participants chose the same scenario for different targets in Studies 2 and 3 (where more than one target was presented). We found that the similarity in which scenarios were chosen for different targets aligned with the psychological distance of the target. That is, when choosing which scenario would most likely elicit a mental state, people most often chose the same option for the self and close targets, while they were most likely to choose different scenarios for the self and far targets (see Supplement S8).

Each trial was followed by jittered fixation drawn from approximate Poisson distributions with intervals equal to the study’s TR and the following means: M_1 = 1.67s, M_2 = 1.53s, M_3 = 1.4 s. Study 1 included 0.25 s of fixation in addition to the jittered fixation time.

2.4. FMRI data acquisition and analysis

2.4.1. Preprocessing—fMRI data from each study were preprocessed similarly (Table 2). Data from all studies underwent coregistration and normalization to the 2 mm isotropic ICBM 152 template with SPM8 (default implementation in Study 1, DARTEL in Studies 2

and 3). FSL's slicetime correction and unwarping were applied in Studies 2 and 3. No explicit smoothing was applied at this stage.

2.4.2. General linear model contrasts—Preprocessed fMRI data were entered into general linear models (GLM) using SPM 8 and the wrapper package SPM8w (https://github.com/ddwagner/SPM8w) in Study 1, and SPM12 via SPM12w (https://github.com/wagner-lab/spm12w) in Studies 2 and 3. Boxcar regressors were created for each condition of interest (i.e., every target-state pairing). Note, while Study 2 trials were grouped by target person, no blocked regressors were included. In Study 2, the similarity matrices produced from the GLM contrasts were corrected for spurious correlations arising from the blocked design (Thornton et al., 2019). There were no other substantial correlations among the regressors of interest (Fig. S5). Trial onsets aligned with the presentation of the prompt, and a trial ended when the participant made a choice or when time maxed out (see Table 1 Scenario Duration). These regressors were convolved with a canonical hemodynamic response function and entered into the GLM along with covariates of no interest, including run means and trends, and six head motion parameters. GLM analyses resulted in one contrast map per target-state pair (Study 1: 1 modality × 1 target × 60 states = 60 maps; Study 2: 1 modality × 3 targets × 25 states = 75 maps; Study 3: 1 modality × 2 targets × 30 states = 60 maps) per participant. These maps represent how participants thought about each target person experiencing each state in each modality condition. These resulting patterns of contrast values were used in all subsequent analyses.

2.4.3. Representational, similarity analysis—Our primary analyses tested the stability of mental state representations in the brain. To do so, we used representational similarity analysis (RSA) and suppressed intercept linear mixed models, as follows. RSA reflects the overall structure of mental state representations by measuring how similar each mental state representation is to every other mental state representation. We can thus use RSA to compare neural representations (Kriegeskorte et al., 2008). Specifically, we can look at the structure people apply for each target, and compare these structures across targets. If people use similar representational structures across different targets, this would mean that they apply similar mental state concepts across targets. We tested whether these structures were significantly similar to each other using suppressed intercept mixed models. These models include a coefficient that reflects the extent of this structural similarity that is not dependent on individual neural patterns being consistent across participants.

All analyses were conducted across the entire brain using a functional whole-brain parcellation (https://identifiers.org/neurovault.collection:2099). The 200 parcels in this functional parcellation were defined based on meta-analytic coactivations in Neurosynth (neurosynth.org; de la Vega et al., 2016). This functional parcellation divides the brain into discrete regions using a data-driven approach to determine which voxels respond similarly throughout past research. As such, we avoid imposing artificial region shapes and sizes via the searchlight approach and reduce the number of multiple comparisons that we need to correct for. This, in turn, increases our statistical power and reduces the computational demand of running mixed models at every point in the brain. For each participant, within each parcel, we generated target-specific representational similarity matrices (RSMs) by

calculating the Pearson correlation between every pair of mental state patterns within a single target (Fig. 2b). This resulted in one 60 × 60 matrix for the one (generic) target in Study 1, three 25 × 25 matrices for the three (self, close, and far) targets in Study 2, and two 30 × 30 matrices for the two (self and far) targets in Study 3, per parcel, per subject. These RSMs are a measurement of how people think about mental state concepts for each target.

Studies 2 and 3 included different subsets of the mental states included in Study 1. To allow for comparison across studies, we expanded all matrices to 60 × 60 with empty rows for mental states that were not shown in that study. This allowed us to calculate the Spearman rank correlation between each of these first-order RSMs (which consists of continuous correlation values) to create a single second-order RSM across all targets, participants, and studies (Fig. 2c; see Fig. S6 for the average second-order RSM). Note that, because mental states were imbalanced across the three datasets, Study 1 may have been driving the structure of these first-order mental state geometries. However, when we exclude Study 1, we found highly congruent results (Fig. S1). The final second-order RSM consisted of 185 rows and columns: one per target per participant per modality/study (20 participants × 1 target × 1 modality + 35 participants × 3 targets × 1 modality + 30 participants × 2 targets × 1 modality = 185).

The first-order RSMs reflect the overall structure, or geometry, of the mental state representational space that a participant held for a particular target in a particular modality. The second-order RSM shows the similarity of these geometries across targets, participants, studies, and modalities. This second-order RSM thus allows us to answer the following questions: Did the same participant think about two different targets' states similarly? Did a participant in Study 2 (lexical stimuli) think about the far target's mental states similarly to how a participant in Study 3 (pictorial stimuli) thought about the far target's mental states?

We conducted three analyses to test if neural representations of mental states remain stable across different (i) targets, (ii) modalities, and (iii) targets and modalities, respectively. Each analysis must be run separately rather than including them all into one model. This is because we use suppressed intercept models, which allows us to model *similarity* (rather than differences) across targets and/or modalities. If we were to combine all three analyses into one, the model would become rank-deficient.

All three stability analyses proceeded similarly. For example, for the target analyses, if people use the same mental state concepts when thinking about different targets, then we should see similar (i.e., correlated) first-order RSMs associated with different targets. Each correlation value between two targets' first-order RSMs corresponds to a cell in the second-order RSM (Fig. 2b–c). Thus, we can test how similarly people think about different targets' mental states by looking at the mean of the cells in the second-order RSM that refer to first-order RSMs from different targets. That is, we measure the average correlation value between first-order RSMs of different targets and test if that value is greater than 0. If so, this suggests a common structure to the mental state spaces (i.e., first-order RSMs) even across different targets.

We implemented this test by applying suppressed intercept linear mixed effects models. In this analysis, we try to explain how similar the representational structure of mental states is based primarily on the target person. To test for this effect of interest–namely, if there is a similar structure across different target people–we included two binary fixed effect predictors for same-target vs. different-target data (Fig. 3a). The same-target predictor has 1 s in cells that compare data from the same target and 0s in all other cells. The different-target predictor is the inverse of the same-target predictor and includes 1 s in cells that compare first-order RSMs from different targets and 0s elsewhere.

Unlike a typical linear (mixed) model, in which the beta reflects the difference between the levels of a categorical variable, our suppressed intercept model (i.e., the intercept is set to 0) coefficients represent the means of each condition. In other words, each beta corresponds to the mean of all cells (in the second-order RSM) that correspond with a 1 in that predictor RSM (rather than the difference between cells with a 1 and cells with a 0; Fig. 3d). Because we are not modeling the intercept, it is necessary to include both of these fixed effects in order to model the within-target similarity and the between-target similarity, and thereby also the overall mean of the data (see Supplement S7; Fig. S4). Thus, in this model, the beta value for the different-target predictor represents the average similarity in the structure of mental state representation between different target people. By running significance testing (ß > 0) on this value, we determined whether or not the overall structure of mental states was similar across different targets in a given parcel. The model also included random intercepts by participant and study, along with a random intercept to account for mental states in the same modality (Fig. 3a).

We tested this model across the entire brain. Specifically, we fit the model in each of the 200 regions of our parcellation (see above for description of parcellation, Section 2.4.3) to map out where mental state representations are similar across different targets. The resulting *p*-values were calculated via Satterthwaite approximation for degrees of freedom, and multiple comparisons across the 200 parcels were controlled via Holm correction.

To test the stability of mental state representations across modalities, we conducted the same type of analysis with slightly different predictors: as fixed effects, we included one same-modality predictor, which indicates whether or not a cell compares data from the same modality (1 = same modality, 0 = different modalities; Fig. 3b). We also included one different-modality predictor that reflects whether or not a cell compares data from different modalities (1 = different modalities, 0 = same modality). As above, both terms are necessary to model both within-modality similarity and between-modality similarity. The beta value for the different modality predictor represents how similarly participants' representational spaces of mental states (i.e., the first-order RSMs) are across studies that used different modalities. We also included random effects to account for mental states within the same target person, participant, and study (Fig. 3b). After fitting the model in every parcel, we performed significance testing (using the same procedure described above) on the different-modality coefficient to determine where representations of mental states are stable across different stimulus modalities (Fig. 3b).

Finally, we tested if there were regions with stable mental state representations across both targets and modalities using the same process. In this suppressed intercept linear mixed effects model, we included the following predictors: as fixed effects, we included one same-target/modality predictor, which indicates whether or not a cell compares data from the same target *or* the same modality (1 = same target or same modality, 0 = different targets and different modalities). We also included one different-target/modality predictor that reflects whether or not a cell compares data from different targets *and* different modalities (1 = different targets and different modalities, 0 = same target or same modality). The beta value for the different-target/modality predictor represents how similarly participants think about mental state concepts across different targets *and* different modalities. We also included random effects to participant and study. After fitting the model in every parcel, we performed the same significance testing described above on the different-target/modality coefficient to determine where representations of mental states are stable across different target people and stimulus modalities.

3. Results

3.1. Stability of neural representations across targets

Do people think about mental states in the same way across different targets, or do people think about each target's mind uniquely? To test this, we compared neural representations of states across different target people. We did so using representational similarity analysis (RSA) and a suppressed intercept linear mixed effects model to quantify the extent to which each parcel represented mental states similarly across target people. Significance testing on this coefficient showed robustly stable representations of mental states across different targets, specifically within regions associated with social cognition and mental state representation. These regions include the ventral medial prefrontal cortex (vMPFC), dorsal medial prefrontal cortex (dMPFC), precuneus, bilateral temporoparietal junction (TPJ), and anterior temporal lobe (ATL; Fig. 4a; Table 3). These regions represent others' mental states in similar ways regardless of who is thought to be experiencing these states. We found convergent results when analyzing Studies 2 and 3 independently (see Supplement S4).

3.2. Stability of neural representations across modalities

Do people think about mental states in the same way whether they see the situation or read about it? To test for supramodal representations of mental states, we compared neural representations of states across modalities. We did so using a suppressed intercept linear mixed effects model with RSA to quantify the extent to which parcels represented mental states similarly across stimulus modalities. Similar to the target stability results, we found that large portions of the social brain network, including MPFC, ATL, and bilateral TPJ, showed robustly stable neural representations of mental states across modalities. These results further suggest that these areas encode generalizable representations of mental states (Fig. 4b; Table 3).

3.3. Stability of neural representations across targets and modalities

Are there brain regions that hold stable representations of mental state concepts across both targets and modalities? Once again, we used a suppressed intercept linear mixed effects

model with RSA to quantify the extent to which parcels represented mental states similarly across target people and stimulus modalities. Once again, we found areas of the MPFC, left dlPFC, and TPJ, as well as the dACC and left insula represented mental states similarly across both contextual changes (Fig. 4c; Table 3).

4. Discussion

Each individual experiences mental states in their own idiosyncratic way. Yet, perceivers are able to accurately understand this huge variety of states across the uniqueness of each individual and context. How do people accomplish this feat? Here we demonstrate that people do so by drawing upon a core model of mental state concepts. Across three studies, we find that people apply the same fundamental understanding of mental states, no matter who is experiencing a particular state or how they take in that state. Together, these findings suggest that people's knowledge of mental states is encoded consistently across people and modalities.

We found robustly stable representations in a subset of the regions implicated in social cognition known as the default mode network (Mars et al., 2012, Mitchell, 2008, Schurz et al., 2020, Van Overwalle and Baetens, 2009). Specifically, the MPFC, ATL, and TPJ showed robust stability in mental state representations across both targets and modalities (Fig. 4; Table 3). All of these parcels are consistent with areas associated with mentalizing processes, suggesting that not only do these parcels support social cognition, they also support generalized mental state concepts across contextual specifics.

That said, the precuneus generalized across targets, but not modalities. Given that the two stimulus modalities we used were words and images, it makes sense that the precuneus, which supports mental imagery processes, might not represent mental states similarly across these different visual modalities. Similarly, a portion of the dorsal MPFC generalized across modalities, but not targets, and encoded target-*specific* mental state concepts (see Supplement S5; Fig. S2). The dMPFC has been shown to respond differently to similar and dissimilar others (Mitchell et al., 2006, Tamir and Mitchell, 2010). The current finding supports the idea that different regions play distinct roles in supporting social cognitive processes, and that the dMPFC in particular, may serve to help individualize targets. This and other regions across the social brain, however, respond similarly across modalities. This finding is particularly striking when considering the vast differences in informational cues that we have for understanding and experiencing our states of mind, compared to someone else's. While we are the ultimate authorities on our own feelings, we must rely on entirely external cues to understand what another person is feeling, which will never provide enough information for us to understand their experience fully. Yet, our results suggest that we use a generalizable model of how mental states relate to one another across diverse modalities. These findings are congruent with past literature showing that our brain represents supramodal representations of emotions (Chikazoe et al., 2014, Peelen et al., 2010, Skerry and Saxe, 2014) and words (Marinkovic et al., 2003). This prior work specifically finds that portions of the MPFC, ATL, and TPJ encode mental states as core concepts, invariant across different contexts. Importantly, the current findings do not preclude the possibility that other regions also have generalized representations of mental state concepts that we could not

detect with the present design, or that regions associated with mental state representation can be modulated by contextual features. Rather, our results suggest that after filtering out the context-specific factors, brain regions spanning the default mode network encode a common core structure of how mental states relate to one another across a diverse range of modalities.

These findings suggest that people use a similar core model of emotions for the self and others. We suspect that people develop this generalizable amodal model of mental state representation through the convergence of knowledge about one's own experiences with observations of others' experiences. For example, if someone grew up in a place where everyone was quick to anger, then their mental state model would place anger in a more prominent node than someone who grew up in a more placid environment (Thornton and Tamir, 2021). That is, the landscapes of others' emotions and the cultural norms in which one is embedded define each individual's context-independent model. This model, in turn, constrains both one's personal emotional landscape, as well as one's perception of others'. Likewise, a person's own experiences can constrain how they perceive others' experiences. People are often egocentric in their social inferences, including their inferences about others' emotions (Trilla et al., 2020). For example, an individual who is often quick to anger might overperceive anger in others. This bidirectional pathway between learning and inference has implications for both emotional contagion and clinical research. In clinical psychology, mood disorders affect how a person thinks about their own feelings (Leppänen, 2006). Since our results suggest that people use the same model to understand others' mental states as their own, mood disorders may impact a person's social perception of others in the same way.

Our results offer robust evidence for generalization across diverse targets and stimuli. However, we note at least three limitations to generalizability. First, our samples of all three datasets were not diverse in age, demographics, education, location, and more. As such, future researchers should practice caution before generalizing these results to other populations. Future research should explore if these results hold in different cultures, especially since recent research suggests large cultural variability in emotion concepts and expression (Jackson et al., 2019). Second, the stimuli used were relatively artificial (short phrases, static images) compared to the rich sources of information used in everyday life. We did not explore other modalities commonly used when making inferences about others' minds, including speech and dynamic stimuli. These other types of information sources may provide further insight into when and why people use this generalizable framework. Finally, this project focused on the neural representations of mental state *concepts*. As such, our results cannot speak to how people think about mental state *experiences*. First-hand experiences of states may well show greater variation than the generalizable mental state concepts discussed here.

5. Conclusion

When interacting with others, we must consider their perspective, thoughts, and feelings. Here, we found that, while context may modulate processing of this information, we likely draw on core features of these mental states that provide a consistent generalizable amodal

model across the varying situations of our daily lives. This shared core can facilitate deeper understanding of each other's internal lives.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Elisa Baek for comments on previous versions of this manuscript.

Funding

This work was supported by the National Institute of Mental Health R01MH114904 (D.I.T); and the National Science Foundation Graduate Research Fellowship Program (M.E.W.).

References

Barrett LF, Mesquita B, Ochsner KN, Gross JJ, 2007. The experience of emotion. Annu. Rev. Psychol 58 (1), 373–403. doi:10.1146/annurev.psych.58.110405.085709. [PubMed: 17002554]

Bedny M, Caramazza A, Grossman E, Pascual-Leone A, Saxe R, 2008. Concepts are more than percepts: the case of action verbs. J. Neurosci 28 (44), 11347–11353. doi:10.1523/JNEUROSCI.3039-08.2008. [PubMed: 18971476]

Chikazoe J, Lee DH, Kriegeskorte N, Anderson AK, 2014. Population coding of affect across stimuli, modalities and individuals. Nat. Neurosci 17 (8), 1114–1122. doi:10.1038/nn.3749. [PubMed: 24952643]

Cikara M, Fiske ST, 2011. Bounded empathy: neural responses to outgroup targets' (mis)fortunes. J. Cognit. Neurosci 23 (12), 3791–3803. doi:10.1162/jocn_a_00069. [PubMed: 21671744]

Claypool HM, Hugenberg K, Housley MK, Mackie DM, 2007. Familiar eyes are smiling: on the role of familiarity in the perception of facial affect. Eur. J. Soc. Psychol 37 (5), 856–866. doi:10.1002/ejsp.422.

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. J. Neurosci 36 (24), 6553–6562. doi:10.1523/jneurosci.4402-15.2016. [PubMed: 27307242]

de Vignemont F, Singer T, 2006. The empathic brain: how, when and why? Trends Cognit. Sci 10 (10), 435–441. doi:10.1016/j.tics.2006.08.008. [PubMed: 16949331]

Epley N, 2008. Solving the (real) other minds problem. Soc. Personal. Psychol. Compass 2 (3), 1455–1474. doi:10.1111/j.1751-9004.2008.00115.x.

Harris LT, Fiske ST, 2011. Dehumanized perception. Z. Psychol 219 (3), 175–181. doi:10.1027/2151-2604/a000065. [PubMed: 24511459]

Haslam N, Loughnan S, 2014. Dehumanization and infrahumanization. Annu. Rev. Psychol 65 (1), 399–423. doi:10.1146/annurev-psych-010213-115045. [PubMed: 23808915]

Jackson JC, Watts J, Henry TR, List J, Forkel R, Mucha PJ, …, Lindquist KA, 2019. Emotion semantics show both cultural variation and universal structure. Science 366 (6472), 1517–1522. doi:10.1126/science.aaw8160. [PubMed: 31857485]

Kriegeskorte N, Mur M, Bandettini P, 2008. Representational similarity analysis – connecting the branches of systems neuroscience. Front. Syst. Neurosci 2 (4), 1–28. doi:10.3389/neuro.06.004.2008. [PubMed: 18958245]

Leppänen JM, 2006. Emotional information processing in mood disorders: a review of behavioral and neuroimaging findings. Curr. Opin. Psychiatry 19 (1), 34–39. doi:10.1097/01.yco.0000191500.46411.00. [PubMed: 16612176]

Marinkovic K, Dhond RP, Dale AM, Glessner M, Carr V, Halgren E, 2003. Spatiotemporal dynamics of modality-specific and supramodal word processing. Neuron 38 (3), 487–497. doi:10.1016/S0896-6273(03)00197-1. [PubMed: 12741994]

Mars RB, Neubert FX, Noonan MAP, Sallet J, Toni I, Rushworth MFS, 2012. On the relationship between the “default mode network” and the “social brain. Front. Hum. Neurosci 6 (189), 1–9. doi:10.3389/fnhum.2012.00189. [PubMed: 22279433]

Meyer ML, Spunt RP, Berkman ET, Taylor SE, Lieberman MD, 2012. Evidence for social working memory from a parametric functional MRI study. Proc. Natl. Acad. Sci 109 (6), 1883–1888. doi:10.1073/pnas.1121077109. [PubMed: 22308468]

Mitchell JP, 2008. Contributions of functional neuroimaging to the study of social cognition. Curr. Direct. Psychol. Sci 17 (2), 142–146. doi:10.1111/j.1467-8721.2008.00564.x.

Mitchell JP, Heatherton TF, Macrae CN, 2002. Distinct neural systems subserve person and object knowledge. Proc. Natl. Acad. Sci 99 (23), 15238–15243. doi:10.1073/pnas.232395699. [PubMed: 12417766]

Mitchell JP, Macrae CN, Banaji MR, 2006. Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. Neuron 50 (4), 655–663. doi:10.1016/j.neuron.2006.03.040. [PubMed: 16701214]

Nisbett RE, Wilson TD, 1977. Telling more than we can know: Verbal reports on mental processes. Psychol. Rev 84 (3), 231–259. doi:10.1037/0033-295X.84.3.231.

Peelen MV, Atkinson AP, Vuilleumier P, 2010. Supramodal representations of perceived emotions in the human brain. J. Neurosci 30 (30), 10127–10134. doi:10.1523/JNEUROSCI.2161-10.2010 . [PubMed: 20668196]

Saxe R, Kanwisher N, 2003. People thinking about thinking peopleThe role of the temporo-parietal junction in “theory of mind. NeuroImage 19 (4), 1835–1842. doi:10.1016/S1053-8119(03)00230-1. [PubMed: 12948738]

Schurz M, Radua J, Tholen MG, Maliske L, Margulies DS, Mars RB, …, Kanske P, 2020. Toward a hierarchical model of social cognition: a neuroimaging meta-analysis and integrative review of empathy and theory of mind. Psychol. Bull doi:10.1037/bul0000303.

Skerry AE, Saxe R, 2014. A common neural code for perceived and inferred emotion. J. Neuroscience 34 (48), 15997–16008. doi:10.1523/JNEUROSCI.1676-14.2014.

Tamir DI, Mitchell JP, 2010. Neural correlates of anchoring-and-adjustment during mentalizing. Proc. Natl. Acad. Sci 107 (24), 10827–10832. doi:10.1073/pnas.1003242107. [PubMed: 20534459]

Tamir DI, Thornton MA, Contreras JM, Mitchell JP, 2016. Neural evidence that three dimensions organize mental state representation: rationality, social impact, and valence. Proc. Natl. Acad. Sci 113 (1), 194–199. doi:10.1073/pnas.1511905112. [PubMed: 26621704]

Theriault J, Waytz A, Heiphetz L, Young L, 2020. Theory of mind network activity is associated with metaethical judgment: an item analysis. Neuropsychologia 143 (January), 107475. doi:10.1016/j.neuropsychologia.2020.107475. [PubMed: 32360298]

Thornton MA, & Tamir DI (2021) (n.d.). The organization of social knowledge is tuned for prediction. In Gilead M & Ochsner KN (Eds.), The Neural Bases of Mentalizing. Springer Press. 10.1007/978-3-030-51890-5_14

Thornton MA, Weaverdyck ME, Mildner JN, Tamir DI, 2019. People represent their own mental states more distinctly than others. Nat. Commun 10 (2117), 1–9. doi:10.1038/s41467-019-10083-6 . [PubMed: 30602773]

Thornton MA, Weaverdyck ME, Tamir DI, 2018. The brain represents people as the mental states they habitually experience. Nat. Commun (2019) 1–10. doi:10.1038/s41467-019-10309-7. [PubMed: 29317637]

Thornton MA, Wolf S, Reilly BJ, Slingerland EG, Tamir DI, 2020. The 3d mind model characterizes how people understand mental states across modern and historical cultures. PsyArXiv doi:10.31234/osf.io/m5p74.

Todorov A, Gobbini MI, Evans KK, Haxby JV, 2007. Spontaneous retrieval of affective person knowledge in face perception. Neuropsychologia 45 (1), 163–173. doi:10.1016/j.neuropsychologia.2006.04.018. [PubMed: 16759672]

Trilla I, Weigand A, Dziobek I, 2020. Affective states influence emotion perception: Evidence for emotional egocentricity. Psychol. Res doi:10.1007/s00426-020-01314-3.

Trope Y, Liberman N, 2010. Construal-level theory of psychological distance. Psychol. Rev 117 (2), 440–463. doi:10.1037/a0018963. [PubMed: 20438233]

Van Overwalle F, Baetens K, 2009. Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. NeuroImage 48 (3), 564–584. doi:10.1016/j.neuroimage.2009.06.009. [PubMed: 19524046]

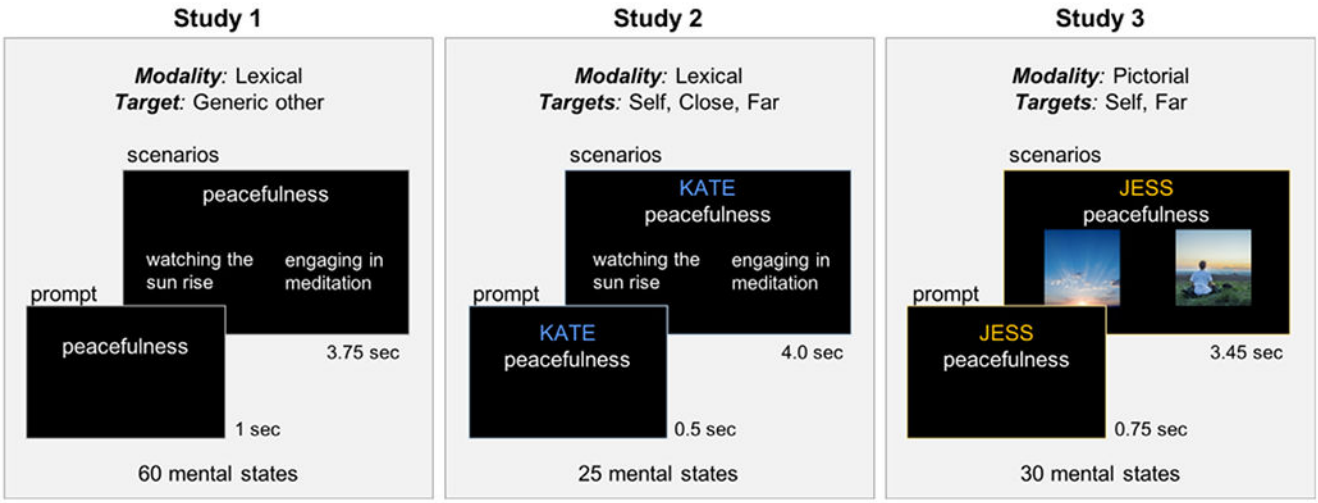

Fig. 1. Sample trial from each study.

In all three studies, participants were shown a prompt with the current mental state (e.g., peacefulness) followed by two scenarios that elicit that mental state. On each trial, participants had to decide which of the two scenarios would most likely elicit the current state in the current target. In Study 1, they chose for a generic other; in Studies 2 and 3, participants chose for a specific target. Self trials were cued with the word “SELF”; socially close and far targets were cued with the name of that person (e.g., “KATE”). Close targets were nominated by the participant, while far targets were created by the experimenter to be unfamiliar and dissimilar to the participant. Scenarios were presented as a short phrase in Studies 1 and 2, and as images in Study 3.

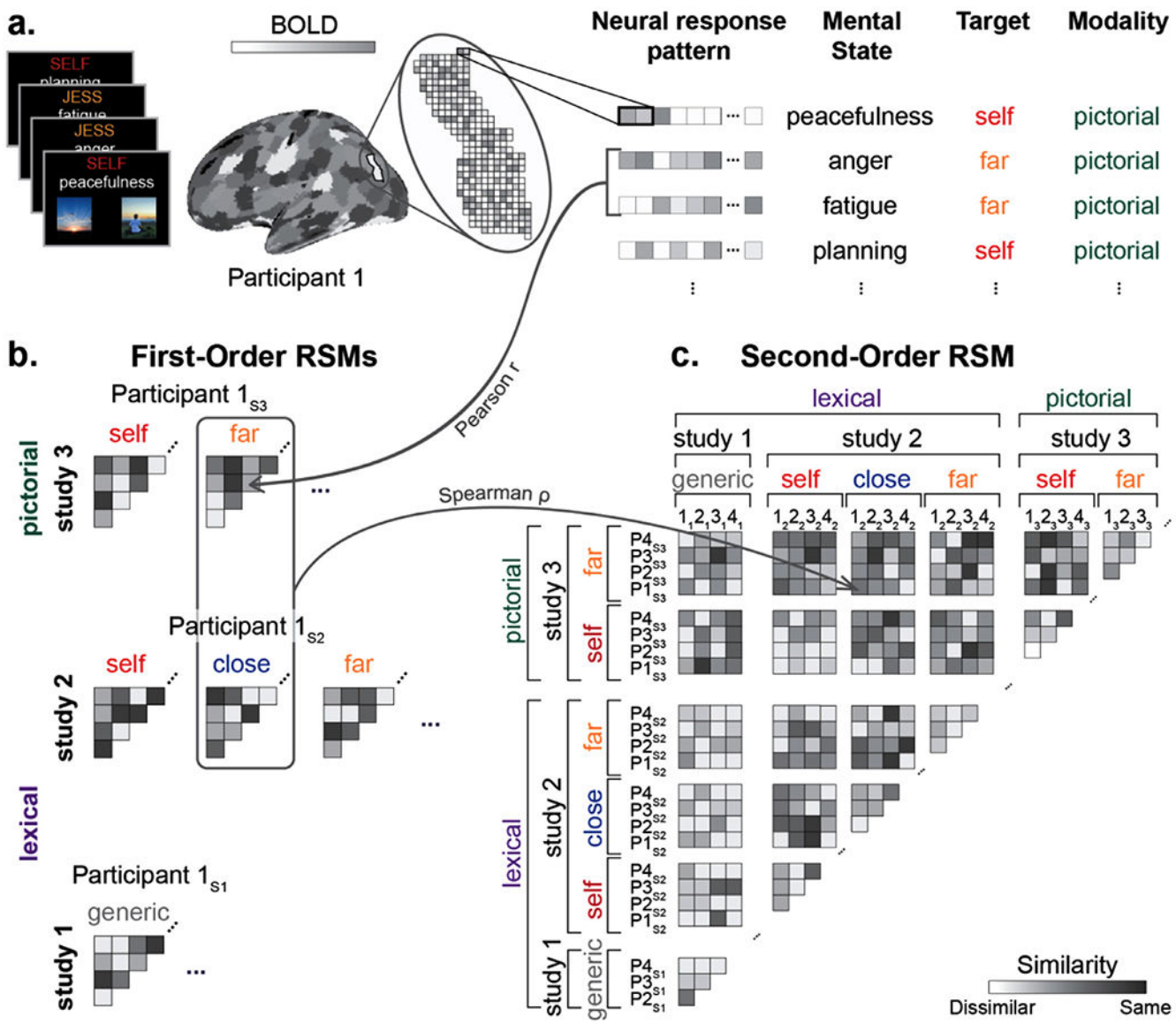

Fig. 2. Generating the second-order representational similarity matrix.

Within each participant and neural parcel, (a) neural response patterns elicited by each target-state pair were extracted and (b) compared via Pearson correlation. These correlation coefficients were then organized into target-specific RSMs. Each of these first-order RSMs was (c) Spearman correlated with every other RSM, within and across studies. These values were arctan transformed and organized into a second-order RSM. This second-order neural RSM was entered into suppressed intercept models (Fig. 3).

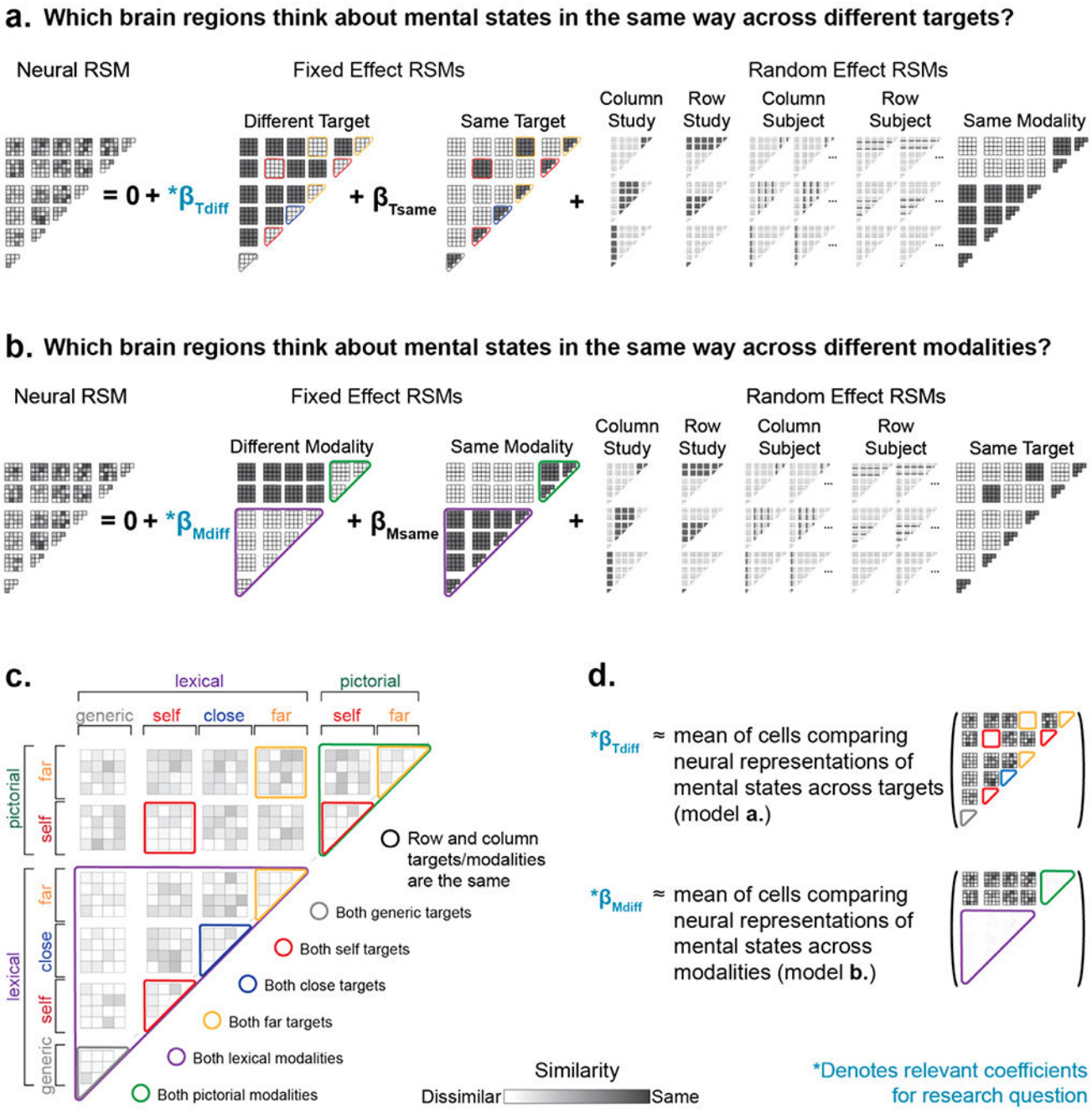

Fig. 3. Statistical models.

In separate suppressed-intercept linear mixed effects models, we tested which brain regions hold stable representations of mental state concepts across (a) targets or (b) modalities. (c) The second-order neural RSM (Fig. 2c) was the dependent variable in each mixed-effects model with individual predictors for same and different targets/modalities, and random effects for subject, study, and modality/target. (d) Because we suppressed the intercept (set equal to 0), the different targets/modalities coefficient (blue with asterisk) reflected the mean correlation values of cells that compare first-order RSMs from different targets/modalities

(dark cells). That is, it measured the average similarity of mental state representations across targets/modalities. If this coefficient was significantly greater than 0, then the relevant brain region showed stable structures of mental state representation across targets/modalities.

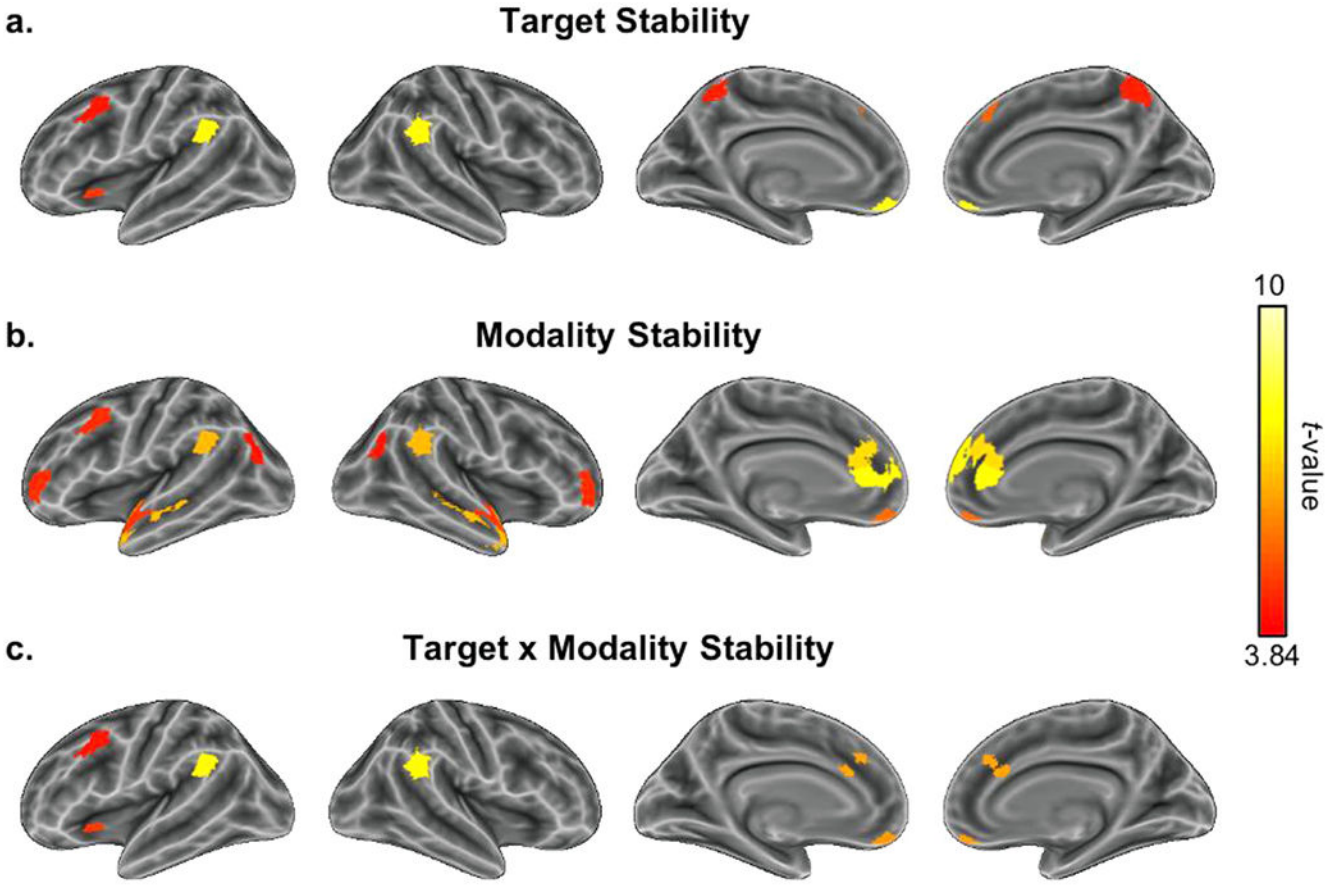

Fig. 4. Stable mental state representations across targets and modalities.

Participants showed stable neural representations of mental states across (a) target people and (b) stimulus modalities in parcels associated with social cognition. A subset of these regions (parts of the MPFC, dlPFC, and bilateral TPJ) showed stable representations across (c) both targets and modalities. In addition, the precuneus showed stable representations across targets. Only significant *t*-values (α = .05) after correcting for multiple comparisons are shown.

Table 1

Experimental design parameters.

Study	Targets	Modality	Number of States	Prompt Duration (s)	Scenario Duration (s)	Trial Order
Study 1	Generic other	Lexical	60	1.00	3.75	Intermixed
Study 2	Self, Close, Far	Lexical	25	0.50	4.00	Blocked
Study 3	Self, Far	Pictorial	30	0.75	3.45	Intermixed

Table 2

MRI acquisition parameters for all three studies.

Study	Scanner	Head Coil	Resolution (mm)	TR (s)	TE (ms)	FA (°)	Runs	TRs
Study 1	Trio	32-channel						
T2*			2.50 × 2.51 × 2.51	2.5	30	90	16	162
T1			1.2 × 1.2 × 1.2	2.2	1.54	7		
Study 2	Prisma	64-channel						
T2*			2 × 2 × 2	2.25	32	70	12	243
T1			1 × 1 × 1	2.3	2.27	8		
Study 3	Prisma	64-channel						
T2*			2 × 2 × 2	1.4	32	70	12	209
T1			1 × 1 × 1	2.3	2.27	8		

Table 3

Parcels that stably represent mental states across targets and modalities.

		Across Targets				Across Modalities				Across Targets and Modalities			
Approx. ROI	**Index**	**ß**	***t***	***p***		**ß**	***t***	***p***		**ß**	***t***	***p***	
Temporo-parietal Junction	195	0.04	9.15	<.001	***	0.05	6.89	<.001	***	0.04	7.19	<.001	***
Ventromedial Prefrontal Cortex	147	0.03	8.11	<.001	***	0.02	5.83	<.001	***	0.02	5.91	<.001	***
Left Dorsolateral Prefrontal Cortex	174	0.02	5.92	<.001	***	0.02	4.69	.003	**	0.02	4.01	.017	*
Anterior Temporal Lobe	47	0.02	6.30	<.001	***	0.02	4.92	.001	***	0.02	3.90	1.000	
Left Anterior Insula	88	0.01	4.55	1.000		0.01	4.19	.060	+	0.01	4.45	0.003	**
Dorsal Anterior Cingulate Cortex	65	0.02	5.03	1.000		0.02	3.82	1.000		0.02	5.97	<.001	***
Dorsomedial Prefrontal Cortex	142	0.04	5.06	1.000		0.04	8.07	<.001	***	0.02	2.53	1.000	
Anterior Cingulate Cortex	148	0.04	5.59	1.000		0.04	8.18	<.001	***	0.03	4.58	1.000	
Dorsal Anterior Cingulate Cortex	116	0.03	3.78	1.000		0.04	7.40	<.001	***	0.03	3.73	1.000	
Temporal Pole	36	0.04	6.29	.519		0.04	7.05	.004	**	0.03	3.01	1.000	
Inferior Parietal Lobule	181	0.05	2.78	1.000		0.03	4.38	.006	**	0.06	3.85	1.000	
Lateral Frontal Pole	145	0.01	2.75	1.000		0.02	4.78	.015	*	0.00	0.59	1.000	
Caudate Nucleus	28	0.00	1.95	1.000		0.01	3.68	.078	+	0.00	1.96	1.000	
Pre-Supplementary Motor Area	154	0.02	5.56	<.001	***	0.03	3.03	1.000		0.02	3.35	1.000	
Precuneus	179	0.01	4.62	.002	**	0.01	2.89	1.000		0.01	2.86	.911	
Left Anterior Insula	132	0.01	3.85	.039	*	0.01	2.65	1.000		0.01	2.83	.996	
Posterior Parietal Cortex	152	0.01	4.07	.019	*	0.01	2.85	1.000		0.00	1.25	1.000	
Insula	180	0.01	3.84	.040	*	0.01	3.33	1.000		0.00	1.35	1.000	
Amygdala	129	0.01	3.68	.073	+	0.01	2.97	1.000		0.00	1.95	1.000	

+ p < .100;

* *p* < .050;

** *p* < .010;

*** *p* < .001

Note. Only regions with a trending or significant result are displayed. All *p*-values are corrected for multiple comparisons using the Holm method