The Neural Organization of Social Knowledge

The Neural Organization of Social Knowledge

A dissertation presented

by

Mark Allen Thornton

To

The *Department of Psychology*

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Psychology

Harvard University

Cambridge, Massachusetts

May 2017

Dissertation Advisor: Professor Jason P. Mitchell Mark Allen Thornton

The Neural Organization of Social Knowledge

Abstract

Humans enjoy social lives of terrific complexity, but this sociality exacts high demands on the individual. The typical adult must interact with hundreds of others on an ongoing basis. Each such person in our life embodies a unique constellation of mental states and traits. How can we possibly navigate this social complexity? Two crucial tools for any navigator are their map and compass: knowledge of the terrain, and a set of cardinal directions they use to organize this knowledge. Across three studies, this dissertation investigates the social analogs of these navigational tools: knowledge of other people, and the psychological dimensions the brain relies on to organize this information. Using advanced neuroimaging techniques, these studies systematically map domains of social knowledge to reveal their organizing principles. Paper 1 investigates the neural representational of others' mental states. Results suggest that three psychological dimensions – rationality, valence, and social impact – explain nearly half of the variance in how the brain makes sense of others' thoughts and feelings. Paper 2 investigates the domain of personally familiar others – how people represent individuals, such as friends and family, with whom they interact on a regular basis. The data suggest these individuals are represented by both coarse and fine-grained patterns distributed throughout brain regions involved in social cognition. Paper 3 investigates the representation of famous others, with the aim of identifying dimensions which explain how the brain represents well-known individuals with whom one lacks direct contact. Results indicate that four established theories of person

perception can each accurately predict patterns of brain activity associated with famous individuals. Moreover, a synthetic model combining these established theories achieved two-thirds of the accuracy of a hypothetical ideal theory. Combining data from Papers 1 and 3 yields another insight: encoding models trained to predict person-specific patterns of brain activity can also predict state-specific activity patterns. This suggests that the brain represents others' mental states and traits in a partially-shared representational space. Together, these Papers advance our understanding of how the brain organizes social knowledge, and contribute to synthesizing more general and accurate theories of social cognition.

Table of Contents

Acknowledgments

I thank my advisor, Jason Mitchell, for guiding me through the course of my doctoral research, and for helping me bring out the exciting and profound sides of my sometimes arcane and technical research. I thank the members of my dissertation committee, Talia Konkle, Jim Haxby, and Fiery Cushman, for their invaluable feedback on this document and the ideas therein. I am also deeply grateful to many other members of the faculty and staff at Harvard University, including Professors Mina Cikara, Mahzarin Banaji, and Leah Somerville, my colleague in teaching, Patrick Mair, and the CBS team, including Tammy Moran and Ross Mair.

I am profoundly indebted to my labmates and collaborators in the Social Cognitive and Affective Neuroscience Lab, including Diana Tamir, Juan Manuel Contreras, Joe Moran, Sara Verosky, Jamil Zaki, Joachim Norberg, Eshin Jolly, Rita Ludwig, and Franchesca Ramirez. I also greatly appreciate the help of all the research assistants who have worked on my projects over the past six years, including Spencer Dunleavy, Abigail Orlando, Ryan Song, Ava Zhang, Annie Rak, Brenda Li, Radhika Rastogi, and Eve Wesson.

Several institutions and funding agencies played important roles in making my doctoral research possible. I am grateful to Harvard University for providing funding for my first and sixth years, and for being a reliable landlord throughout my time in Cambridge. I thank the National Science Foundation for the Graduate Research Fellowship (DGE1144152) which supported me during my second, third, and fourth years. I also thank the Sackler Scholar Programme in Psychobiology, which supported me in my fifth year. Additionally, I thank the Center for Open Science and the Harvard Dataverse for providing venues to freely share the data and code from my dissertation research with the world.

I would not have made it through the last six years without the friendship and support of many fellow students in the Harvard Psychology Department. Thank you to my cohort, and particularly to the Doctors of Velocity – Steven Felix, Molly Dillon, Gus Cooney, Bria Long, and Ken Allen – who have accompanied me throughout this journey. Thank you also to my neighbors and teatime buddies on WJH 15, and my officemate Benedek Kurdi. Thank you to the many graduate students who passed through Psych 1950 during the three years I TF'd the course – you pushed me to be a better teacher, and made me proud with the work you've done since.

Finally, I wish to express my deep and abiding gratitude to my parents – Roy Thornton and Susan Rodriguez. They raised me to be a person of integrity and compassion, inspired me to pursue scientific research, and provided me with constant moral support throughout my studies.

Introduction

How does one mind make sense of another? This question resides near the center of human experience. On an interpersonal level, bridging the gap between minds provides us with the friendships, rivalries, and intimacies that create meaning in our lives. However, social connections do not just make life worth living, they also make living possible. Whether they live in tiny bands or global societies, humans rely on other humans to survive. The unique achievements of human civilization are, at least in part, predicated on the ability of people to understand and cooperate with one another. Indeed, considerable evidence suggests that the advantages of social understanding have exerted a major influence on the evolution of the human brain (Dunbar, 1998; Herrmann, Call, Hernandez-Lloreda, Hare, & Tomasello, 2007; Holekamp, 2007; Moll & Tomasello, 2007; Reader & Laland, 2002; van Schaik & Burkart, 2011).

The question of how people understand one another is not just important at humanistic or evolutionary levels of analysis. It also presents a profound computational problem, because the information processing system being understood is roughly as complex as the system doing the understanding. Moreover, each person possesses a mind of enormous complexity. In modern life, humans must interact with a wide array of stimuli – from apples to operating systems – but the minds of other humans remain arguably the most complicated. The challenge is further magnified by the fact that we cannot devote our lives to understanding a single mind, but must instead achieve a passable understanding of the minds of hundreds of other individuals on an ongoing basis (McCormick, Salganik, & Zheng, 2010).

If we do not perceive the task of understanding other people as so profoundly challenging, this may be because evolution has endowed us with specialized mechanisms for understanding one another. Over the past decades, researchers in the burgeoning field of social

cognitive neuroscience have identified the neural hardware that supports theory of mind – or mentalizing – the ability to understand others' minds. Around the turn of the millennium, the default assumption of many psychologists was that social and nonsocial processes of a similar nature would be subserved by shared neural substrates (Blakemore, Winston, & Frith, 2004). However, shortly thereafter, a number of landmark studies revealed that a network of brain regions – now commonly referred to as the social brain network, or "social brain" for short – are selectively engaged by social versus nonsocial cognition (J. P. Mitchell, Heatherton, & Macrae, 2002; J. P. Mitchell, Macrae, & Banaji, 2004; J. P. Mitchell, Macrae, & Banaji, 2005; Rilling, Sanfey, Aronson, Nystrom, & Cohen, 2004; Saxe & Kanwisher, 2003; Saxe & Wexler, 2005). Brain regions typically implicated in this network including both dorsal and ventral medial prefrontal cortex (MPFC), the anterior temporal lobe (ATL), the superior temporal sulcus (STS) extending back to the temporoparietal junction (TPJ), and the medial parietal cortex (MPC), including the precuneus, posterior cingulate, and retrosplenial cortex. The surprising dissociation between these regions and those more heavily involved in analogous nonsocial cognition – such as the lateral prefrontal cortex (LPFC) and inferior temporal (IT) cortex – was one of the first major contributions of social cognitive neuroscience to social psychology (J. P. Mitchell, 2008).

In the years since, the robustness of this distinction has become clear, with hundreds of functional magnetic resonance imaging (fMRI) studies supporting the results of these first few (for reviews, see Amodio & Frith, 2006; Van Overwalle & Baetens, 2009). However, neuroimaging has not just brought new distinctions to social cognition, but also new unions. For instance, the social brain network was found to overlap considerably with the "default mode" or default network (Mars et al., 2012; Schilbach, Eickhoff, Rotarska-Jagiela, Fink, & Vogeley, 2008), a set of brain regions more active at baseline resting state than during many active tasks

(Buckner, Andrews-Hanna, & Schacter, 2008; Raichle et al., 2001), leading to speculation that social cognition is the brain's default process. Commentators also observed that social brain engagement appears to unite a broad range of social, affective, and self-referential processes, suggesting that the various domains traditionally explored by social psychologists may have a deep underlying connection (J. P. Mitchell, 2009).

Considerable progress has been made in understanding processes within the social brain network and its various nodes. For example, a ventral-dorsal distinction in MPFC has repeatedly been implicated in distinct neural processes for mentalizing about different types of target people: similar versus dissimilar (J. P. Mitchell, Macrae, & Banaji, 2006), cooperative versus competitive (Decety, Jackson, Sommerville, Chaminade, & Meltzoff, 2004; Gallagher, Jack, Roepstorff, & Frith, 2002), close versus far (Krienen, Tu, & Buckner, 2010), and familiar versus unfamiliar (Welborn & Lieberman, 2015). Ventral MPFC has been implicated in self-representation (Kelley et al., 2002; J. P. Mitchell, Banaji, & Macrae, 2005; Northoff et al., 2006), in using the self as a model for others (Decety & Grèzes, 2006; Tamir & Mitchell, 2010), and in trading off utility for self and other (Hare, Camerer, Knoepfle, O'Doherty, & Rangel, 2010; Zaki, López, & Mitchell, 2014). The ATL has been implicated in identity representation (Anzellotti, Fairhall, & Caramazza, 2013; Gorno-Tempini & Price, 2001; Kriegeskorte, Formisano, Sorger, & Goebel, 2007) and social knowledge (Ross & Olson, 2009). The computational goal of the social brain network as a whole has been linked to self-projection (Buckner & Carroll, 2007; Spreng, Mar, & Kim, 2009) – the ability to displace one's perspective along various psychological dimensions into different people, places, and times (Parkinson, Liu, & Wheatley, 2014; Tamir & Mitchell, 2011).

Although much ground has been gained in understanding information *processing* in the social brain, less has been gained in understanding social information *representation* or *organization*. In this respect, we are like an aspiring chef with an understanding of how to manipulate food – chopping, crushing, mixing, frying, baking, and so forth – but little knowledge of the landscape of possible ingredients or their essential properties. To put the problem in more computational terms, we have started to come to grips with the functions and operators of the social mind – and their neural hardware – but we have made less progress in understanding the implementation of the variables – the objects, classes, and values over which those functions operate. Fortunately, innovative analytic approaches may now allow us to redress this imbalance by learning how the social brain organizes its contents.

Since their introduction (Haxby et al., 2001; Kriegeskorte & Bandettini, 2007; Kriegeskorte, Goebel, & Bandettini, 2006; Kriegeskorte, Mur, & Bandettini, 2008), advanced neuroimaging analysis methods under the umbrella of multivoxel (or multivariate) pattern analysis (MVPA) have revolutionized the study of how the brain represents content (for reviews, see Haxby, 2012; Norman, Polyn, Detre, & Haxby, 2006). These methods have led to a renaissance in how we understand the cognitive organization of domains such as visual objects (Cohen, Alvarez, Nakayama, & Konkle, 2016; Khaligh-Razavi & Kriegeskorte, 2014; Kriegeskorte, Mur, Ruff, et al., 2008) and scenes (Park, Konkle, & Oliva, 2015), biological classes (Connolly et al., 2012), nouns (Just, Cherkassky, Aryal, & Mitchell, 2010; T. M. Mitchell et al., 2008), and other semantic classes (Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016; Huth, Nishimoto, Vu, & Gallant, 2012). Recently these approaches have started to appear in the study of social cognition as well (Hassabis et al., 2014; Leshinskaya, Contreras, Caramazza, &

Mitchell, 2017; Skerry & Saxe, 2015; Stolier & Freeman, 2016), but our understanding of the neural organization of social knowledge is, as-yet, nascent.

The overarching goal of this dissertation is to explore how people understand each other by investigating how the brain organizes representations of other minds. The three papers presented here attempt to map how the brain represents three domains of social knowledge, and thereby discover the primary organizing principles the mind uses to make sense of other people. In Paper 1, I investigate the neural organization of mental state representation, in an effort to understand how we make sense of others' thoughts and feelings. In Paper 2, I turn to the domain of personally familiar others, probing which brain regions and psychological dimensions support our ability to richly simulate the minds of people close to us. Finally, in Paper 3, I test which of several prominent theories of person perception can explain the informational basis of mentalizing about people we do not know personally (in this case, famous individuals).

Each of these papers follows a similar approach. First, I carefully select stimuli from the domain in question – mental states, personally familiar people, or famous people – and determine the positions of these stimuli on the dimensions of relevant existing or synthetic psychological theories. These selection and placement procedures rely on a variety of techniques including behavioral judgments, online ratings, web scraping and text analysis, and sophisticated optimization algorithms. Second, I use fMRI to measure patterns of neural activity elicited in participants' brains by mentalizing about the states or people selected. These mentalizing tasks are presented lexically, but involve making everyday inferences about, or simulating the minds of, other people. Finally, using a combination of pattern analyses – representational similarity analysis (Kriegeskorte, Mur, & Bandettini, 2008), searchlights (Kriegeskorte et al., 2006), and voxelwise encoding models (Huth et al., 2012; T. M. Mitchell et al., 2008) – I test which regions

of the brain represent stimuli from the domain in question, and which psychological dimensions or theories best explain the corresponding patterns within these regions.

Across all three papers, I observe remarkably consistent results. Rich information about states and people is distributed in regions across the entire social brain, at multiple spatial scales from individual voxels to the entire network. Existing psychological theories prove able to explain much of the variance in the patterns of activity elicited within these regions, particularly with respect to states (Paper 1) and famous people (Paper 3). In both of these cases, we observe that synthesizing existing theories yields a similar three dimensional structure that achieves roughly half to two-thirds of the predictive accuracy expected from a hypothetical ideal theory. Moreover, encoding models prove capable of accurate prediction across these two data sets – accurately reconstructing state-specific patterns after being trained on person-specific patterns – suggesting that the brain represents others' mental states and traits in a partially shared representational space. Together, these results advance our understanding of how the brain organizes knowledge of other people, and thereby, how we each bridge the gap between minds.

Paper 1: Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence

Tamir, D. I.*, Thornton, M. A.*, Contreras, J. M., & Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences, 113*(1), 194-199.
*equal contributions

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

One crucial tool for any navigator is a compass: a set of dimensions that help organize the contents of the world. By attending to the position of others' mental states on key dimensions, humans might reduce the complexity of others' minds to just a few essential elements – coordinates on a map. Might navigators of the world of mental states make use of such an intuitive compass? Research in other domains of cognition suggests this might be possible: the brain has a demonstrated capacity for extracting and capitalizing on useful regularities in the world. For example, our object representation system makes use of dimensions such as size and animacy to organize its processing tracts (Konkle & Caramazza, 2013). Here we explore the possibility that similar principles may organize our representations of other people's minds.

Decades of research in social cognitive neuroscience, primarily using functional magnetic resonance imaging (fMRI), have already implicated a well-defined set of brain regions in the process of thinking about mental states: thinking about the lives and minds of others reliably engages a network including medial prefrontal cortex (MPFC), medial parietal cortex (MPC), temporoparietal junction (TPJ), superior temporal sulcus (STS), and the anterior temporal lobe (ATL; for review, see J. P. Mitchell, 2008; Van Overwalle & Baetens, 2009). However, this relatively young field has yet to explain *how* the social brain's hardware processes the richness and complexity of others' mental states. Fortunately, research in psychology supplies a set of theories regarding how people might organize their knowledge of mental states. The dimensions of these theories include valence and arousal (Posner, Russell, & Peterson, 2005; Russell, 1980), warmth and competence (Cuddy, Fiske, & Glick, 2008; Fiske, Cuddy, Glick, & Xu, 2002), agency and experience (Gray, Gray, & Wegner, 2007), emotion and reason, mind and body (Forstmann & Burgmer, 2015), social and nonsocial (Britton et al., 2006; J. P. Mitchell, 2008, 2009), and uniquely human and shared with animals (Haslam, 2006). Any of these dimensions might plausibly play a role in organizing our understanding of mental states. But which, if any, do we spontaneously employ during mentalizing? If a dimension actually matters to the way people typically think about others' mental states, we should see evidence that the brain organizes its activity around that dimension. However, merely locating *where* in the brain mental state processing occurs—as social neuroscience has done so well already—cannot tell us *how* these regions represent mental states.

Fortunately, new analytic techniques in functional neuroimaging, under the umbrella of multivariate or multivoxel pattern analysis (MVPA), enable us to bridge these levels of analysis. MVPA examines activity in distributed sets of voxels, allowing for discrimination between

stimuli by their associated patterns of activity even when absolute magnitudes of activity remain constant. In this study, we use the form of MVPA known as representational similarity analysis (Kriegeskorte, Mur, & Bandettini, 2008) to test which psychological dimensions organize people's understanding of mental states. This analyses works by measuring the extent to which neural patterns of activity can be predicted from theories of representational organization. To illustrate, the dimension *arousal* would predict that "ecstasy" and "rage" are represented very similarly in the brain, as both are similarly intense mental states. In contrast, the dimension *valence* would predict that "ecstasy" and "rage" are represented very differently in the brain, as one state is very positive, whereas the other is very negative. Both predictions can be tested by measuring the extent to which patterns of neural activity elicited by thinking about a person in ecstasy are similar to those elicited by thinking about a person in a fit of rage. Each dimension makes thousands of predictions about the similarity of each mental state compared to each other mental state; representational similarity analysis allows us to assess the accuracy of all of these predictions simultaneously. Thus we can test which psychological dimensions capture the way the brain encodes others' mental states.

Methods

Neuroimaging data from this study have been deposited at the Harvard University Dataverse (http://dx.doi.org/10.7910/DVN/ELLLZM). Behavioral data and presentation and analysis code are available on the Open Science Framework (https://osf.io/3qn47/). Shared data have been stripped of identifying information.

Participants. A Monte Carlo simulation was used to determine participant and trial numbers consistent with adequate statistical power. We simulated a behavioral similarity matrix and a neural similarity matrix that were correlated at the estimated population effect size of r =

.15. This effect size was thought to be reasonable based on previous work (Kriegeskorte, Mur, & Bandettini, 2008). To generate the behavioral similarity matrix, we simulated activity in a single searchlight. That searchlight consisted of 200 voxels, and 60 separate patterns of activity, to represent each of the mental states. The simulated "activity" within the voxels should be normally distributed ($M = 0$, $SD = 1$) as an approximation for the T-values used in the actual analysis. The 60 x 60 correlation matrix produced by this searchlight was considered the behavioral model. We created patterns of neural activity within the simulated searchlight by taking the 200 voxel by 60 state matrix used to generate the behavioral model and adding additional random noise ~ $N(0, 2.4)$. When these neural patterns were correlated with one another, the resulting neural similarity matrix consistently correlated with the corresponding behavioral matrix at approximately $r = .15$. Since this neural pattern matrix reflects experiment level data, we added additional noise ~ $N(0, 10)$ to represent data from individual trials.

On each iteration of the simulation, a particular participant number and trial (per mental state) were set. Trial-wise neural patterns of searchlight activity were generated for each participant and averaged to produce a single pattern for each participant. These were then converted to similarity matrices and correlated with the overall behavioral similarity matrix. The resulting r values were R-to-z transformed and entered into a t-test across simulated participants. The result of this t-test was tabulated to estimate power. Participant numbers between 2 and 30 and item numbers between 2 and 20 were simulated, with 100 simulation iterations at each combination of these parameters. These simulations indicated that 20 participants with 16 trials per mental state should be adequate to ensure 95% voxelwise statistical power at an uncorrected threshold of $p < .001$.

Participants (N =20) were recruited via the Harvard University Study Pool (16 female; mean age = 22.7 years; range: 18-27 years). All participants were right-handed, native speakers of English, reported no history of neurological problems, and had normal or corrected-to-normal vision. Participants provided informed consent in a manner approved by the Committee on the Use of Human Subjects at Harvard University.

Experimental Design. Participants underwent functional neuroimaging while considering another person experiencing a variety of mental states. The task elicited patterns of neural activity that reflect the representation of each state. On each trial, participants considered 1 of 60 mental states (Table S1). At the onset of the trial, one mental state term was presented for 1 s. This word remained on screen while two very brief scenarios associated with that mental state appeared for 3.75 s, one on the lower left side of the screen, and one on the lower right side. Participants were instructed to report which of the two scenarios they thought would better evoke the mental state in another person. Participants indicated their response using a button box in their left hand by pressing either the middle finger for the left scenario or their index finger for the right scenario. There were no correct answers since both scenarios were pretested to elicit the mental state in question. Each trial was followed by a minimum 250 ms fixation and a randomized jittered fixation period (mean 1.67 s, range 0 – 10 s, in 2.5 s increments). During scanning, participants saw each of the 60 mental states on 16 occasions. Each state was presented once per run over the course of 16 consecutive runs of 405 s each. Participants judged a unique pair of scenarios on each trial; each of 16 scenarios was used only twice over the course of the experiment. Stimuli were presented with PsychoPy (Peirce, 2007).

The 60 mental states in this study were selected to maximize observable differences based on survey ratings from a separate set of participants. To accomplish this, participants

assessed how representative each of 166 mental states was of each of 16 univariate dimensions. Participants (N = 1205) were recruited through Amazon Mechanical Turk and the Harvard University Study Pool to complete one or more of eight online surveys: Emotion/Reason (N = 145), Mind/Body (N = 140), Agency/Experience (N = 145), Warmth/Competence (N = 157), High/Low Arousal (N = 151), Social/Nonsocial (N = 137), Positive/Negative (N = 168), and Shared/Uniquely Human (N = 153). In each survey, participants were provided with definitions of the two dimensions of interest, and then were asked on each trial whether a particular mental state could be categorized along one, both, or neither of the two dimensions of interest in that survey. Across all participants we could thus assess the proportion of trials in which each mental state was (or was not) associated with each dimension. This resulted in continuous ratings between 0 and 1 for each of the 166 mental states on each of 16 psychological dimensions. We used data from the ratings of the 166 mental states along the 16 nominal dimensions to select the optimal set of states. To do so, we ran the resulting 166 x 16 matrix of data through an optimization selection process which iteratively selected a random subset of mental states (separately for subsets between 50 and 98), calculated how well they sampled each dimension using a Kolmogorov-Smirnov test (compared to a uniform distribution), calculated the redundancy of each dimension using Tolerance, and then over 10,000,000 iterations, selected the solution which maximized the former and minimized the later. The optimal solution of 60 mental states was used in the current study.

Table 1.1 *60 mental states used in imaging experiment.*

affection	disgust	intrigue	relaxation
agitation	distrust	judgment	satisfaction
alarm	dominance	laziness	self-consciousness
anticipation	drunkenness	lethargy	self-pity
attention	contemplation	lust	seriousness
awareness	earnestness	nervousness	skepticism
awe	ecstasy	objectivity	sleepiness
belief	embarrassment	opinion	stupor
cognition	exaltation	patience	subordination
consciousness	exhaustion	peacefulness	thought
craziness	fatigue	pensiveness	trance
curiosity	friendliness	pity	transcendence
decision	imagination	planning	uneasiness
desire	insanity	playfulness	weariness
disarray	inspiration	reason	worry

Many of the theories under consideration made similar predictions about mental state representations. We pared down the information contained in the extant models using PCA. The PCA was conducted with respect to the 16 rating dimensions described above and the 60 mental states selected for the experiment. Varimax rotation was used to maximize the interpretability of the factors while maintaining their orthogonality. Parallel analysis (Horn, 1965) and Very Simple Structure (Revelle & Rocklin, 1979) criteria were used to determine component number, with both indicating four factor solutions.

To ensure that the mental state selection process did not bias the factors derived from principal components analysis, an identical analysis was carried out with respect to the 106 mental states not included in the imaging experiment. Very Simple Structure indicated a four factor solution to this analysis for the 106 mental states not included in this study as well. Factor order was not identical across the two solutions, but when rearranged, including reflection where necessary, the factor loadings were reproduced with the following respective reliabilities (*r*s):

.97, .96, .84, .95. The reliability of the solutions not only suggests that the 60 state model did not produce a biased factor structure, but also provides additional evidence for the importance of the identified factors.

The full set of mental states was also used to determine whether an orthogonal PCA rotation was appropriate. The 16 psychological dimensions were subjected to PCA across all 166 rated mental states, with four components retained. These components were then allowed to correlate with each other via an oblique direct oblimin rotation. The resulting factor correlation matrix indicated little tendency for the components to correlate. The highest correlation was between Social Impact and Valence (r = .27), and the mean (absolute value) of the inter-component correlations was r = .12. This suggests that the orthogonal varimax rotation and its concomitant interpretational simplicity may be retained without substantially distorting the relationship between components.

The scenarios presented to subjects in this study were all written to be concise (fewer than 5 words), believable (e.g., “finding $5 on the sidewalk” rather than “winning the lottery”), devoid of personal pronouns, in the present tense, and maximally associated with their respective mental state. To select an optimal set of scenarios, a separate set of participants (N = 795) were recruited through Amazon Mechanical Turk and the Harvard University study pool to complete an online survey that assessed how well each mental state was associated with each scenario. On each trial, participants saw one of the 60 mental states selected using the procedure described above and one of 36 scenarios specific to that mental state. Their task was to rate the degree to which the mental state was associated with the scenario on a scale from 1 (mildly) to 5 (highly). Each participant was presented with 180 such items.

The sets of 16 scenarios for each mental state used in this study were selected (out of a larger set of 36 for each state) by a custom genetic algorithm using participant ratings. Genetic algorithms are optimization programs intended to achieve a desired result by mimicking the mechanisms of organic evolution by natural selection. The algorithm was initiated by randomly generating 100 "strains" with strain defined as any set of 16 scenarios for each mental state. On each of 10,000 iterations the "fitness" of each strain was evaluated (as described below) and strains were selected for reproduction proportional to their fitness raised to the power of 100 (to increase selection pressure) through stochastic universal sampling. Imitating sexual reproduction, two-parent two-child crossover of scenarios within mental state was used to generate a new generation of strains. In this process, each scenario in a "child" strain had an equal probability of being drawn from either of the two "parent" strains. During reproduction each scenario also had .001 probability of "mutating" to a different scenario within the same mental state (even if that scenario appeared in neither of the parent strains). Additionally, the best strain from each generation persisted unchanged to avoid discarding a potential optimum solution.

The "fitness" in this algorithm was determined by four equally weight parts: 1) the scenarios should maximally evoke the mental state of interest – to ensure the appropriateness of each scenario to the mental state in question; 2) they should minimize variability in how well different mental states are evoked – to make sure we were not left with many good examples of one state and bad examples of another; 3) they should minimize variability in how variably scenarios evoke mental states across scenarios – to ensure that choices were not easier for one state (with high variability between scenarios) than another (with lower variability); 4) they should minimize average character length variability across mental states – to ensure that low

level features such as size on retina did not differ across states. The strain with the best fitness at the end of the simulation dictated the scenarios ultimately used the in experiment.

Functional Imaging Procedure. Functional data were acquired using a gradient-echo echo-planar pulse sequence with parallel imaging and prospective motion correction (repetition time = 2500 ms; TE = 30 ms; flip angle = 90°) on a 3T Siemens Trio with standard 32 channel headcoil. Images were acquired using 43 axial, interleaved slices with a thickness of 2.5 mm and 2.51 x 2.51mm in-plane resolution (field of view = 216 mm^2, matrix size = 86 x 86 voxels, 162 measurements per run). Functional images were preprocessed and analyzed with SPM8 (Wellcome Department of Cognitive Neurology, London, UK), using SPM8w. Data were first spatially realigned to correct for head movement and then normalized to a standard anatomical space (2 mm isotropic voxels) based on the ICBM 152 brain template (Montreal Neurological Institute).

A general linear model (GLM) was used to generate participant-specific patterns of activity for each mental state. The model included one regressor for each of the mental states, for a total of 60 regressors of interest. Events were modeled using a canonical hemodynamic response function and covariates of no interest (temporal and dispersion derivatives, session mean, run mean, linear trends, outlier time points and six motion realignment parameters). Boxcar regressors for events began at the onset of the presentation of the mental state. GLM analyses resulted in 60 *t*-value maps, one for each mental state, for each participant. In essence, these maps embody the average neural representation of each state.

We compared neural representations at each voxel in the brain using a searchlight procedure (Kriegeskorte et al., 2006). Patterns of activity for each of the 60 mental states were extracted from participant's GLM-derived *t*-value maps using a spherical searchlight with 4-

voxel radius (~9 mm). To compare the similarity of activity patterns for different mental states, we computed the Pearson correlation between each pair of patterns. Thus two mental states that elicited highly correlated patterns of activity across the searchlight were considered to be more similar to each other. This searchlight procedure resulted in neural similarity matrices at each point in the brain: 60 x 60 matrices whose elements correspond to the correlations between the patterns of neural activity within that searchlight.

We used these estimates of neural similarity to test whether mental states were represented in a manner predicted by the four PCA-derived dimensions. To do so, we made similarity predictions for each dimension with respect to each pair of mental states by taking the absolute difference in their scores on the dimension in question. Multiple regression was used to determine how well the predictions of the PCA-derived dimensions accounted for neural similarity. These regressions generated four maps of unstandardized regression coefficients for each participant, one for each component. The participant-specific maps were smoothed (Gaussian 6 mm FWHM kernel) and entered into random effects analysis using one-sample *t*-tests. The four resulting t-value maps indicate regions of the brain in which differences in the neural patterns elicited by mental states correspond to the differences between mental states along each component. Results were corrected for multiple comparisons via a Monte Carlo simulation using the AFNI (Cox, 1996) 3dClustSim script (estimates of actual smoothness obtained from the four PCA maps and averaged; wholebrain mask from the contrasts constrained voxel number). This indicated that with an uncorrected threshold $p < .001$, a 76 voxel extent was sufficient to yield a corrected threshold of $p < .05$. For visualization, statistical maps were rendered on the cortical surface using Connectome Workbench (Marcus et al., 2011).

To supplement the MVPA approach, we also carried out wholebrain univariate analyses with respect to the four PCA-derived dimensions. To maximize the parallelism between these analyses and the MVPA, we used the same contrast maps produced by the MVPA GLM (i.e. one pattern of betas for each of the 60 mental states for each participant). These maps were smoothed with a 6 mm FWHM Gaussian kernel and then entered into a multiple regression analysis. At each voxel, activity levels from the 60 mental states were regressed onto the scores of the four PCA-derived dimensions. Wholebrain regression maps were combined across participants via voxelwise t-tests. The same voxelwise and cluster correction thresholds were retained for consistency with the multivariate results.

To test whether relevant patterns of activity were represented in a more distributed manner, we conducted an additional network-wide similarity analysis. In this analysis, we generated a single neural similarity matrix per participant based on the pattern of activity across an independently defined network of neural regions. This network was defined using a whole brain omnibus repeated-measures ANOVA across the 60 mental states and 20 participants, which selected any voxels that showed different levels of activity across mental states. Due to the sensitivity of this analysis, voxels were selected at a conservative voxelwise threshold of $p <$ .0001. Note that while this feature selection relies on the same data subjected to MPVA, it is independent of any of the dimensions being tested and thus does not yield biased results. Results of a split-half pattern reliability searchlight analysis yielded highly similar results to those of the ANOVA, suggesting that sensitivity to distributed patterns is not sacrificed by relying on the simpler univariate method.

As with the searchlight analysis, in the network analysis patterns of neural activity were extracted from the entirety of the feature selected area for each of the 60 mental states. These

patterns were correlated to produce a single neural similarity matrix for each participant. These were then averaged to produce a single group-level matrix. The group neural similarity matrix was Pearson-correlated with the similarity matrices generated from each of the four latent dimensions. To generate confidence intervals for these correlations, this procedure was repeated 10,000 times with group similarity matrices based on bootstrapped samples of the 20 participants.

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

To control the scenarios more closely, an automated text analysis was performed to assess several mid-level linguistic properties. In particular, we aimed to control for the concreteness, complexity, and familiarity of the scenarios associated with each mental state. Concreteness and complexity norms were taken from large rating sets (Brysbaert, Warriner, & Kuperman, 2014; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012) with the later measured using age of acquisition as a proxy. Familiarity was based on the SUBTLEXus word frequency measure. For a given scenario, each word's concreteness, complexity, and familiarity values were determined and then averaged to produce a single measure. These were then averaged across scenarios to provide a single score along each linguistic dimension for each of the 60 mental states. The network-level representational similarity analysis was repeated after

partialling out the influence of the linguistic variables on neural pattern similarity. In other words, each linguistic variable was converted to a set of similarity predictions by taking absolute differences and neural pattern similarity was then regressed onto these predictions. These residuals of this regression were then correlated with the four dimensions derived from psychological theories.

Nonmetric multidimensional scaling was used to estimate the proportion of variance in representational space of mental states that could be explained by the three significant PCA-derived dimensions from the representational similarity analysis. We took this approach to better assess the proportion of variance our new model explains in the true dimensions underlying mental state representation. The raw R^2 between the similarity predictions of the PCA-derived dimensions and neural similarity estimates would systematically underestimate this quantity by a quadratic factor, leading us to employ this MDS approach. A 5-D scaling yielded an acceptable stress (.13) below the conventional threshold of .15. Unfortunately, due to the arbitrary orientation of MDS solutions and the high dimensionality of this particular solution, it is beyond the scope of this paper to explore the nature of the unidentified dimensions. However, we do present the results of a 2-D scaling solution to allow the reader to explore the neural similarity more directly. The relative importance of the 5 dimensions were assessed by regressing the original dissimilarity matrix onto similarity predictions made by the 5 MDS dimensions and partitioning the resulting R^2. These estimates were normalized by the total R^2 of the regression and later used as weights. The 5-D scaling dimensions were then individually regressed onto the three significant PCA-based dimensions from the network analysis. The resulting R^2 values were summed into a final estimate of total R^2, with weights based upon the relative importance of each dimension to overall neural similarity as calculated above.

As alternative to null hypothesis significance testing, a cross-validated model selection procedure was also used. This technique proceeded as follows. First, a set of dimensions consisting of between one and four of the PCA-derived dimensions was selected. This step was repeated exhaustively to ultimately include all possible unique combinations of PCs (15 in total). Next, the regression coefficients were simultaneously estimated for all of the selected dimensions with respect to the neural data using non-negative linear least squares. Squared similarity values were used for both behavioral and neural data to allow for later comparison with models containing non-orthogonal dimensions (distances in Euclidean spaces do not sum unless dimensions are orthogonal, but their squares do regardless). The regression coefficients estimated using this approach with the 19 'training' participants were used to weight the squared behavioral similarity values from each dimension. These values were then summed to form a single predictor for the neural similarity space. This predictor was then correlated with the squared neural similarity values of the left out participant to produce a measure of predictive performance. This process was repeated leaving out each participant in turn, and the 20 correlations for each combination of PCs were averaged to produce an overall measure of model performance for that combination. The crossvalidated performance of all possible PC combinations could then be directly compared.

Results

Refining psychological theories. Ratings across many of the 16 theoretical dimensions were highly correlated (Figure 1.1). We distilled the overlapping intuitions embodied in the

original dimensions down to a smaller set of non-redundant dimensions using principal component analysis (PCA).

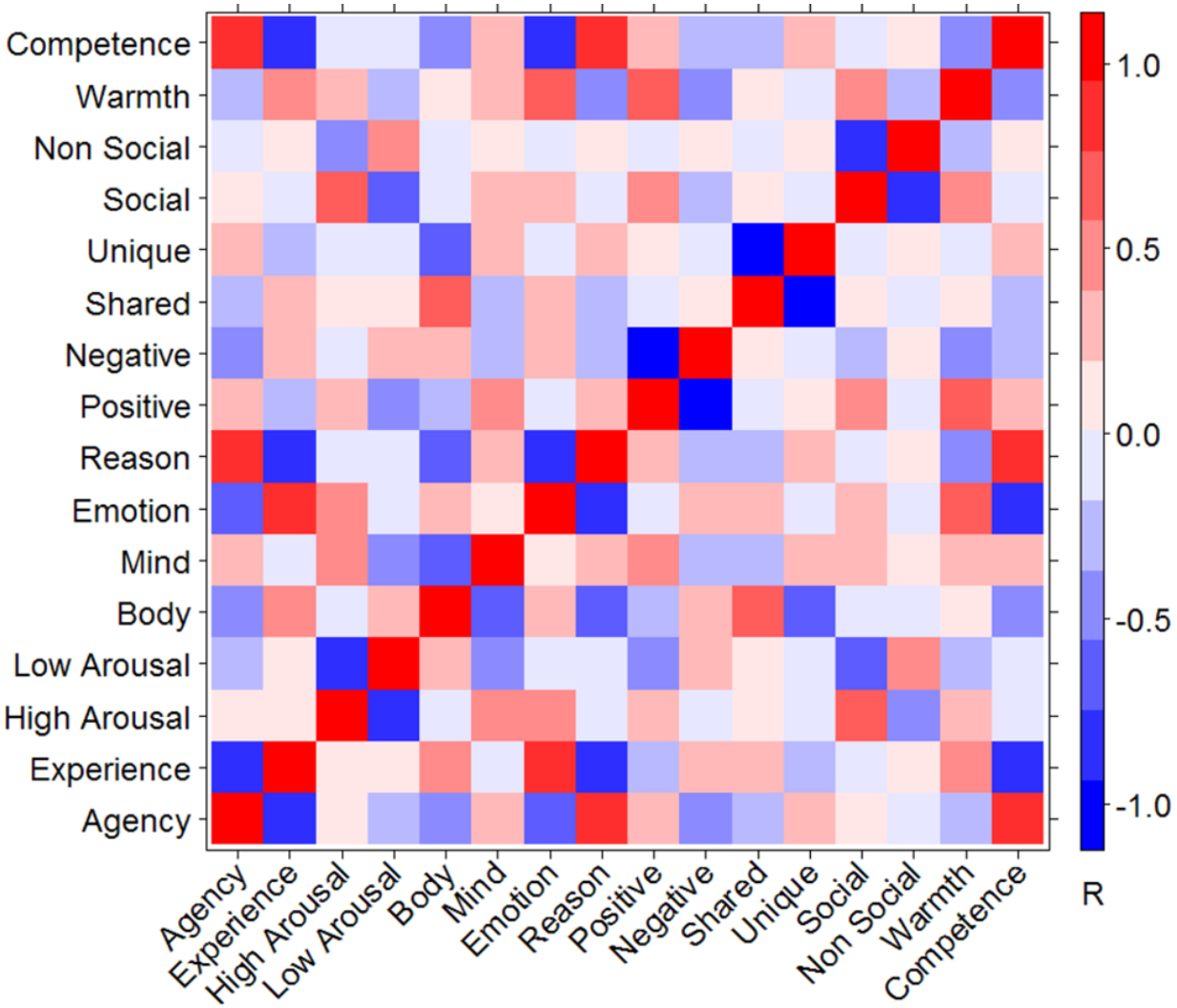


Figure 1.1 *Correlations between theoretical dimensions.* Pearson product-moment correlations between participant ratings of 60 mental states on 16 potential dimensions of mental state representation derived from the existing psychological literature (N = 1,205).

The PCA revealed a much simpler set of four orthogonal dimensions, each with easily interpretable loadings (Figure 1.2). The first component, which we term "rationality," loaded highly in one direction on the original dimensions experience, emotion, and warmth, and loaded

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

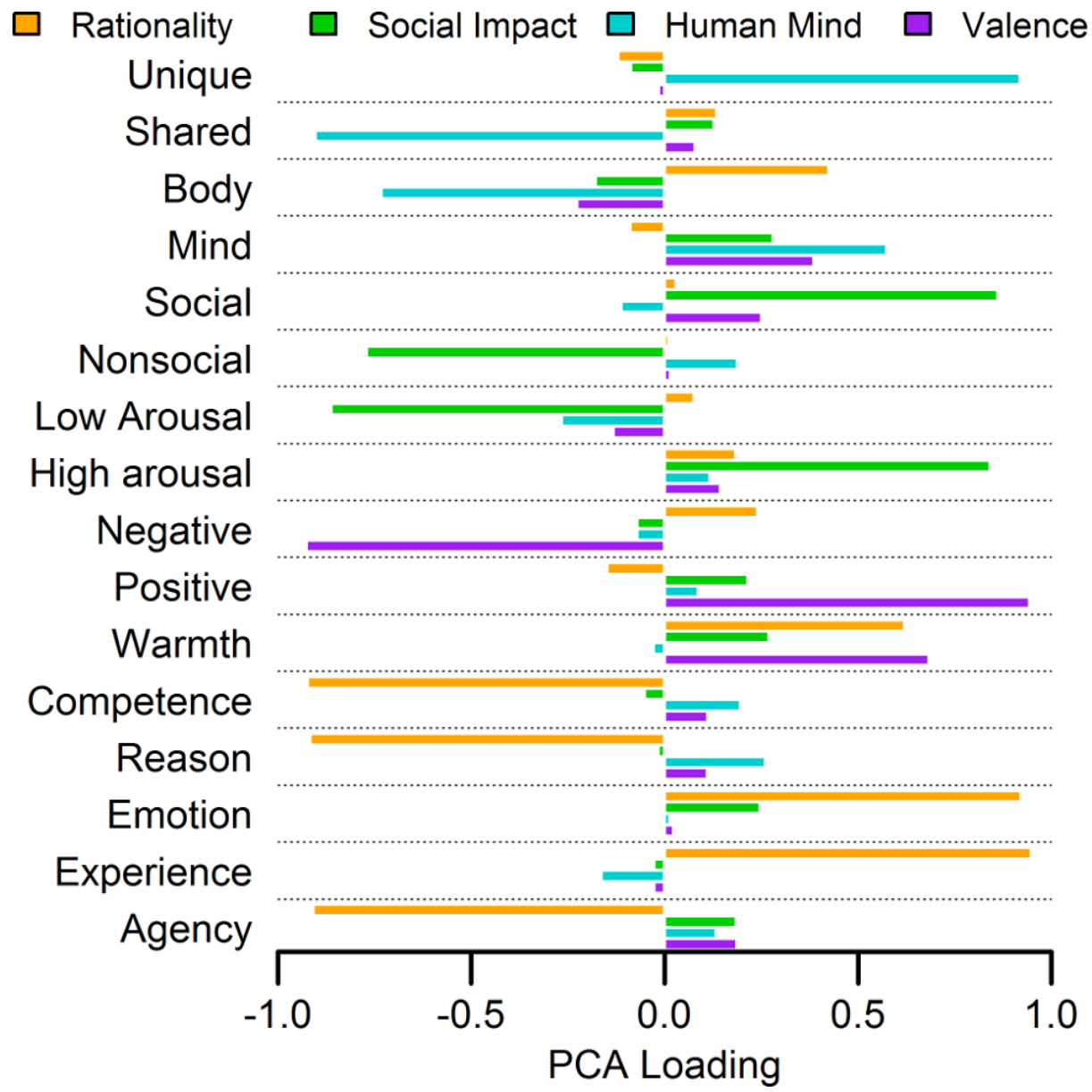


Figure 1.2 *Principal component loadings.* Principal component loadings of the 16 existing theoretical dimensions onto the optimal four dimensional solution.

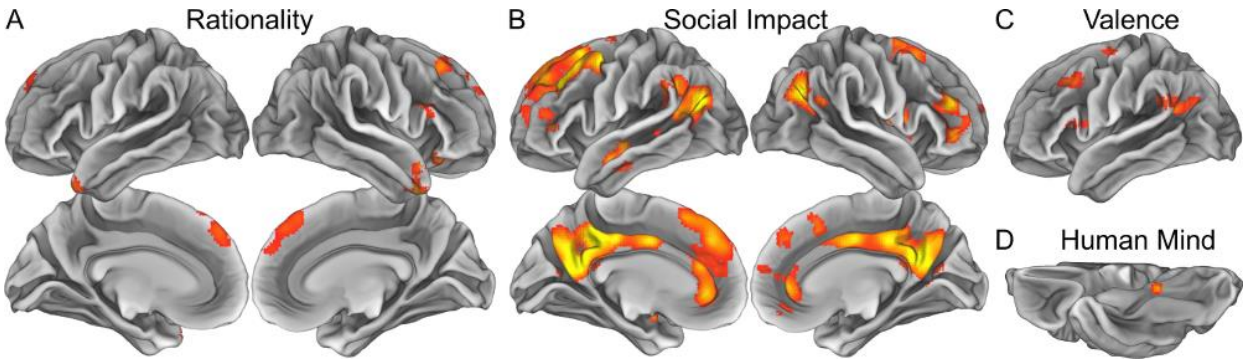


Figure 1.3 *Searchlight results for PCA-derived dimensions.* Within the yellow/orange regions, the similarity of patterns elicited by thinking about mental states can be explained in terms of the corresponding social cognitive dimension extracted from existing theories via PCA ($p < .05$, corrected). Representational similarity searchlight analyses were conducted on each participant and combined through one-sample random-effects *t*-tests.

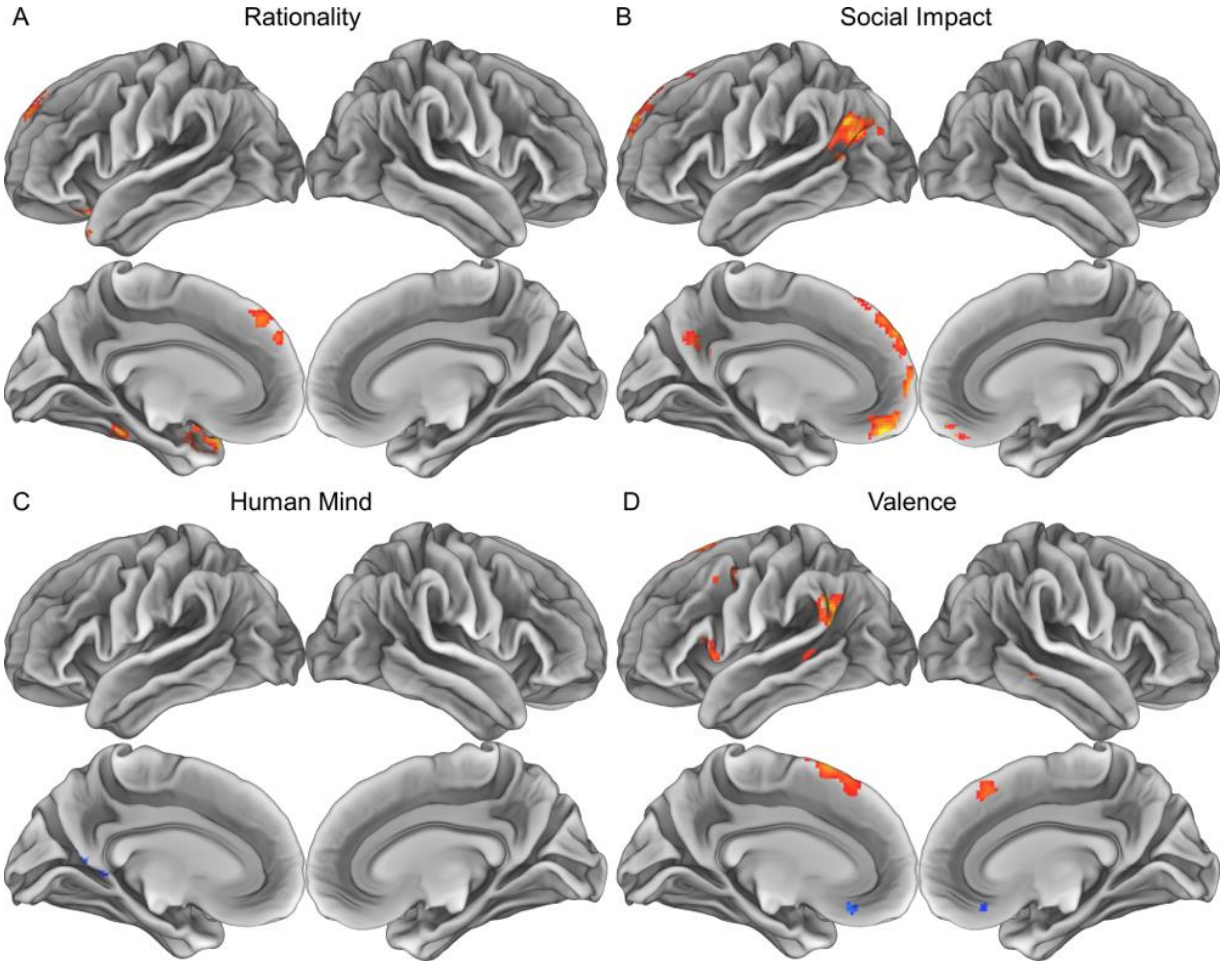


Figure 1.4 *Univariate effects of PCA-derived dimensions*. Significant associations between each of the four PCA-derived dimensions and voxelwise univariate brain activity. Orange voxels indicate activity associated with greater emotionality (and less rationality) of mental states (A), greater social impact (B), or greater negativity (D). Blue voxels indicate activity associated with more shared/bodily states (C), or more positive states (D). Statistical maps resulted from random effects one-sample t-tests across participants, and were statistically corrected using the same standard applied to analogous the multivariate searchlight analysis.

Neural patterns representing PCA dimensions. Representational similarity analysis revealed that three PCA-derived psychological dimensions organize the way the brain represents mental states. Most regions implicated in mental state representation fell within a network of regions previously implicated in social cognition (Figure 1.3 and Table 1.2). The "rationality" dimension predicted the similarity of patterns of neural activity in portions of dorsolateral prefrontal cortex (DLPFC), ventral lateral prefrontal cortex (VLPFC), dorsal medial prefrontal cortex (DMPFC), lateral orbitofrontal cortex (OFC), and the anterior temporal lobe (ATL) bilaterally (Figure 1.3A). The "social impact" dimension robustly predicted neural pattern similarity in a widespread set of regions, including significant clusters in DLPFC, VLPFC, DMPFC, VMPFC, anterior cingulate cortex (ACC), posterior cingulate cortex (PCC), precuneus, temporoparietal junction (TPJ) extending into the posterior superior temporal sulcus (pSTS) and ATL (Figure 1.3B). The "valence" dimension predicted neural pattern similarity in a completely left lateralized set of regions including DLPFC, VLPFC, and TPJ (Figure 1.3C). Valence information was confined to the left lateralized brain regions. Finally, the "human mind" dimension captured a spatially restricted set of neural patterns, predicting representations in only a single region in posterior parahippocampal cortex (Figure 1.3D).

Table 1.2 *Regions representing PCA-derived dimensions*

Dimension/Anatomical Label	x	y	z	Max T	Volume
Rationality					
Anterior temporal lobe	-40	21	-40	4.83	123
Anterior temporal lobe/orbitofrontal cortex	46	13	-36	5.23	584
Ventrolateral prefrontal cortex	48	21	22	4.82	250
Dorsomedial prefrontal cortex/dorsolateral prefrontal cortex	32	33	48	5.13	1449
Social Impact					
Anterior temporal lobe/insula	-38	-3	-12	6.08	890
Posterior cingulate/dorsolateral prefrontal cortex/dorsomedial prefrontal cortex	-8	-57	20	8.29	14306
Anterior temporal lobe	40	41	22	5.88	1106
Insula	48	3	10	8.52	433
Temporoparietal junction	-40	-69	22	6.62	2038
Posterior superior temporal sulcus	-54	-37	0	4.28	78
Temporoparietal junction	46	-63	26	6.40	939
Dorsolateral prefrontal cortex	30	11	60	5.73	955
Human Mind					
parahippocampal gyrus	18	-41	-14	5.29	108
Valence					
Ventrolateral prefrontal cortex	-52	17	10	3.98	99
Temporoparietal junction	-48	-47	28	4.66	646
Dorsolateral prefrontal cortex	-34	23	38	5.01	591
Precentral gyrus	-32	-7	60	4.58	387

Coordinates refer to the Montreal Neurological Institute stereotaxic space.

This analysis identified regions of the brain within which local patterns of activity were predicted by the PCA-based models. To test whether relevant patterns of activity were represented in a more distributed manner, we conducted a network-wide analysis. In this analysis, we extracted a single set of activity patterns from across the entirety of a neural network sensitive to mental state content. As with the wholebrain analysis, the neural similarity of each pair of mental states was estimated and the results correlated with the predictions of the PCA-derived dimensions. Results showed that three dimensions significantly predicted network-

level patterns: “rationality” (r = .16; 95% bootstrap CI [.06, .20]), “social impact” (r = .21; 95% bootstrap CI [.12, .26]), and “valence” (r = .12; 95% bootstrap CI [.04, .17]). The “human mind” dimension (r = .05; 95% bootstrap CI [-.01, .10]) did not (Figure 1.5B).

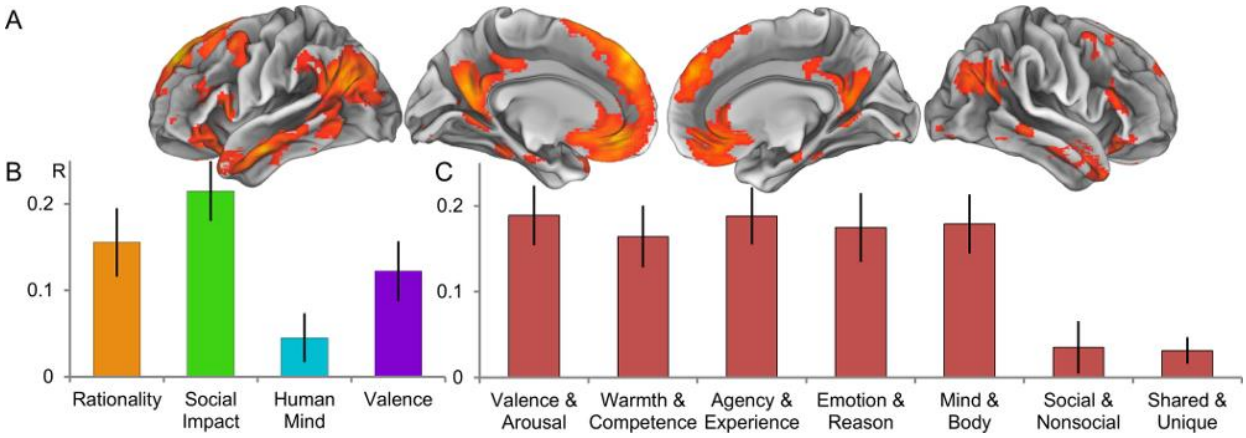


Figure 1.5 *Network-wide representational similarity analysis.* (A) Wholebrain ANOVA used for feature selection (voxelwise $p < .0001$). Different mental states reliably elicited different levels of univariate activity within these regions. (B) Bar graphs of model fits for dimensions derived via principal components analysis from existing psychological theories. (C) Bar graphs of model fits for existing psychological models. All model fits are given in terms of Pearson product-moment correlations between neural pattern similarity and model predictions, with error bars indicating bootstrapped standard errors. Note that bars in (B) refer to individual dimensions derived via PCA whereas bars in (D) indicate the performance of full multidimensional theories. The theoretical advantage of the synthetic model presented here can thus be seen by comparing any one bar in (C) with the combination of the three significant bars in (B).

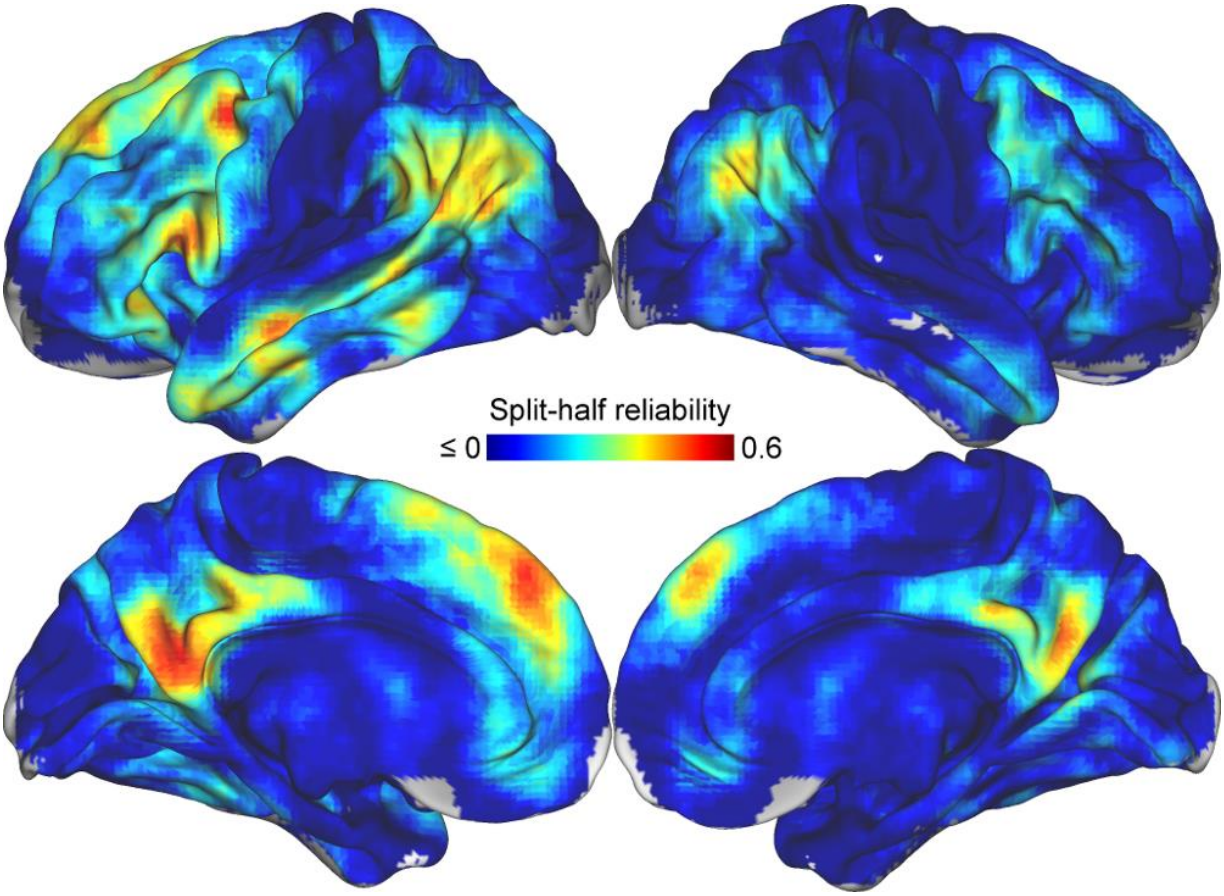


Figure 1.6 *Reliability of similarity searchlights.* The reliability of the neural representations of other's mental states throughout the brain, calculated as the split-half correlation between pattern similarity estimates. Many regions typically implicated in theory of mind demonstrate relatively high reliability, and the results closely mirror the feature selection ANOVA (Figure 1.5A)

MDS allowed us to estimate that the dimensions of rationality, social impact, and valence collectively account for approximately one-third of the variance in neural patterns underlying mental state representation (weighted total R^2 = .33). Disattenuating this value by the dividing it by the reliability of the neural similarity (α = .69) yields a final R^2 = .48.

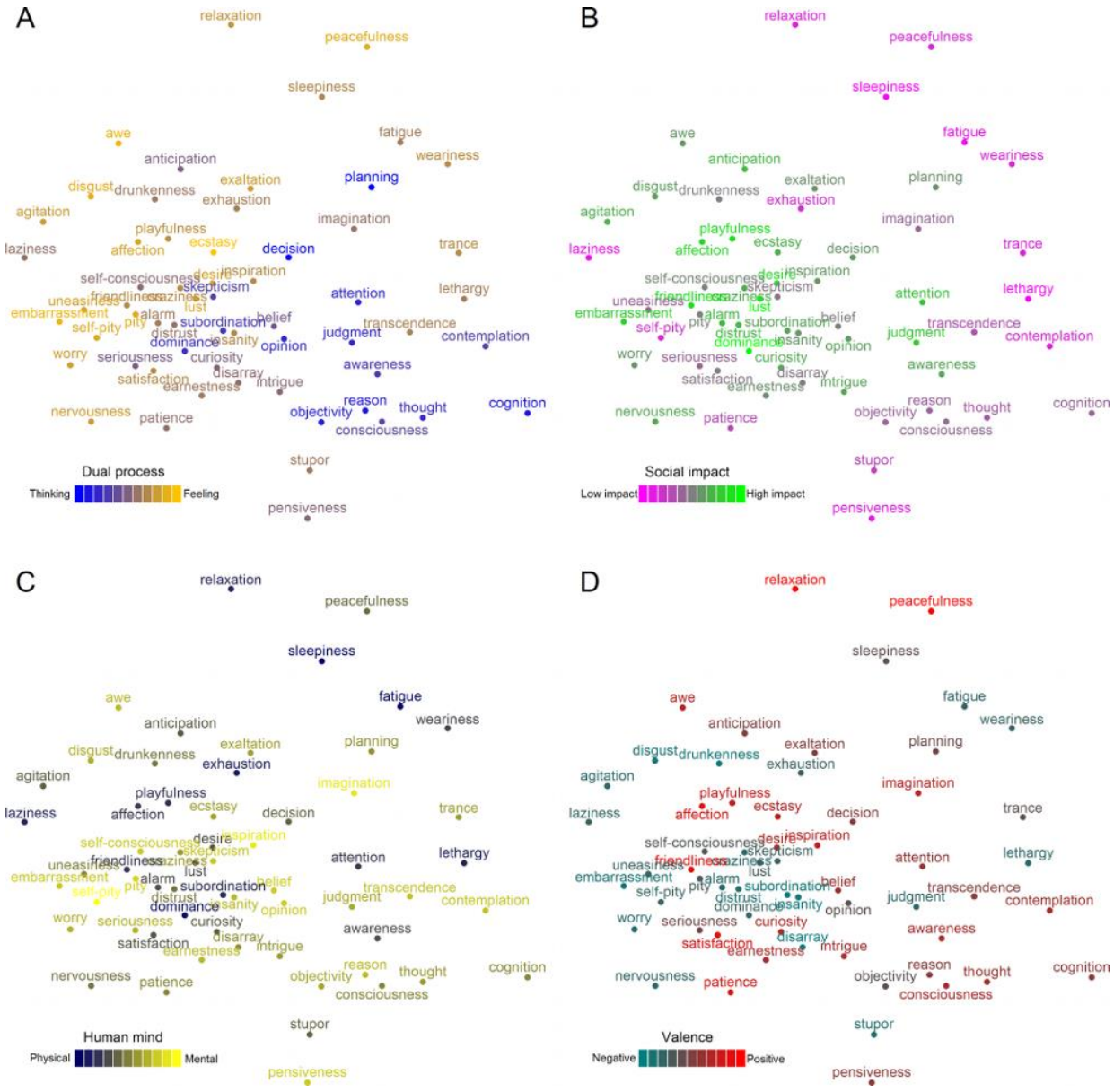


Figure 1.7 *Multidimensional scaling of network-level neural similarity.* Proximity between points indicates greater neural pattern similarity within the social brain network. The same 2-D scaling is presented in A-D, overlaid with each of the four hypothetical dimensions of mental state representation. The 2-D scaling is insufficient to fully capture the differences between patterns elicited for each mental state, but associations between neural space and psychological dimension are still visible.

The results of the network analyses were highly robust to different analytic approaches. Statistically controlling for the influence of scenario concreteness, complexity, and familiarity did not produce any qualitative changes in the outcomes in the RSA results: rationality (r = .15), social impact (r = .21), human mind (r = .05), and valence (r = .12). Moreover, the statistical significance (or lack thereof) at $p < .05$ of the PCA-derived dimensions remained unchanged. This suggests that the influence of these linguistic features cannot account for the predictive power of the psychological dimensions. Results were largely unchanged when using independent. Using independent components analysis (ICA) instead of PCA to generate dimensions, conducting the analysis with Spearman rank correlations, and using a meta-analysis based feature selection method all produced very similar results. Adjusting for changes in order, and ignoring arbitrary sign reflections, ICA and PCA components expressed the following correlations for rationality, social impact, human mind, and valence respectively: rs = .86, .95, .94, and .89. RSA results were similar to those in the primary analysis: rationality (r = .10), social impact (r = .23), human mind (r = .03), and valence (r = .18). The statistical significance of these values did not differ from that reported for the PCA dimensions. Spearman correlation RSA results were also similar to those in the Pearson correlation analysis: rationality (r = .13), social impact (r = .22), human mind (r = .04), and valence (r = .12). Results were also reproducible using independent feature selection (the NeuroSynth reverse inference map at q = .01 for "theory [of] mind"): rationality (r = .19; 95% bootstrap CI [.08, .22]), social impact (r = .20; 95% bootstrap CI [.10, .24]), human mind (r = .02; 95% bootstrap CI [-.03, .08]), and "valence" (r = .10; 95% bootstrap CI [.02, .15]). The fact that these values have not uniformly decreased serves as further evidence for the unbiased nature of the ANOVA feature selection technique.

Inclusion of two-way interactions between the PCA dimensions did not change their statistical significance. However, three of the interaction terms did emerge as significant: rationality x human mind (b = -.003, 95% bootstrap CI [-.004, -.002]), social impact x human mind (b = -.002, 95% bootstrap CI [-.004, -.001]), and social impact x valence (b = .001, 95% bootstrap CI [.0005, .002]).

To assess where the PC dimensions were failing, we calculated the average residual for each mental state and correlated these with three significant PCs. The rationality of a mental state did not predict whether its pattern was chronically predicted to be more or less different to that of other states (r = -.03). The pattern dissimilarity between negative states tended to be slightly overestimated (r = .18). Finally, pattern dissimilarity between highly socially impactful states tended to be substantially underestimated (r = -.66).

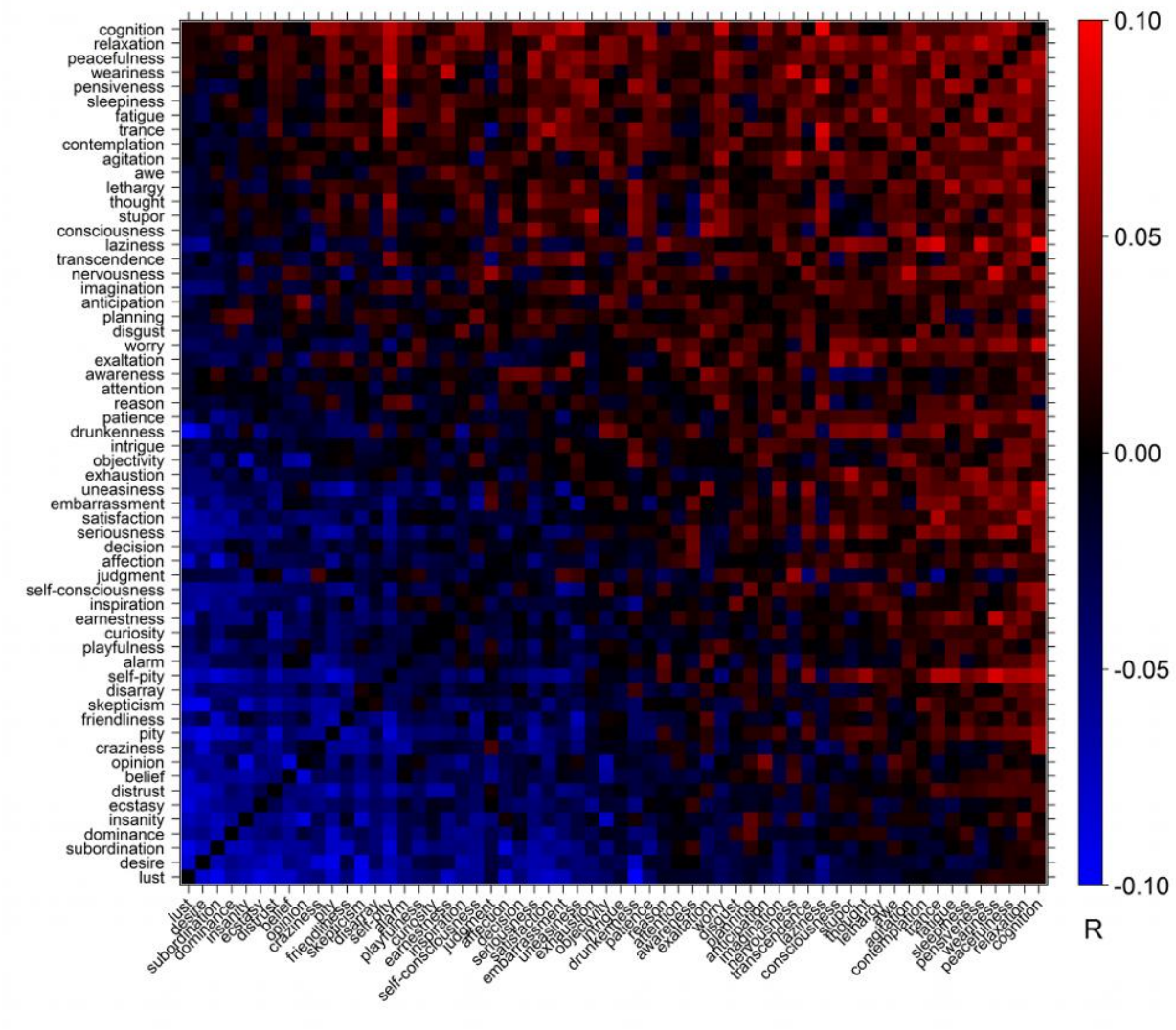


Figure 1.8 *Residual representational dissimilarity matrix.* High positive residuals (red) indicate that mental states were more dissimilar than three significant PCA-derived dimensions would predict. High negative residuals (blue) indicate pairs of mental states that were less different than the PCA-derived dimensions would predict.

Further, results were not contingent on the use of statistical significance: the same three dimensions emerged from a model selection technique based on crossvalidation performance (Khaligh-Razavi & Kriegeskorte, 2014). The 3 dimensional model achieved a crossvalidated r =

.12, relative to a noise ceiling of $r = .28$, again suggesting performance slightly less than half of that of an ideal theory (Figure 1.9). For comparison, the original theoretical models achieved the following performance: valence and arousal ($r = .07$), warmth and competence ($r = .06$), agency and experience ($r = .07$), emotion and reason ($r = .07$), mind and body ($r = .06$), social and nonsocial ($r = .02$), and shared and unique ($r = .01$).

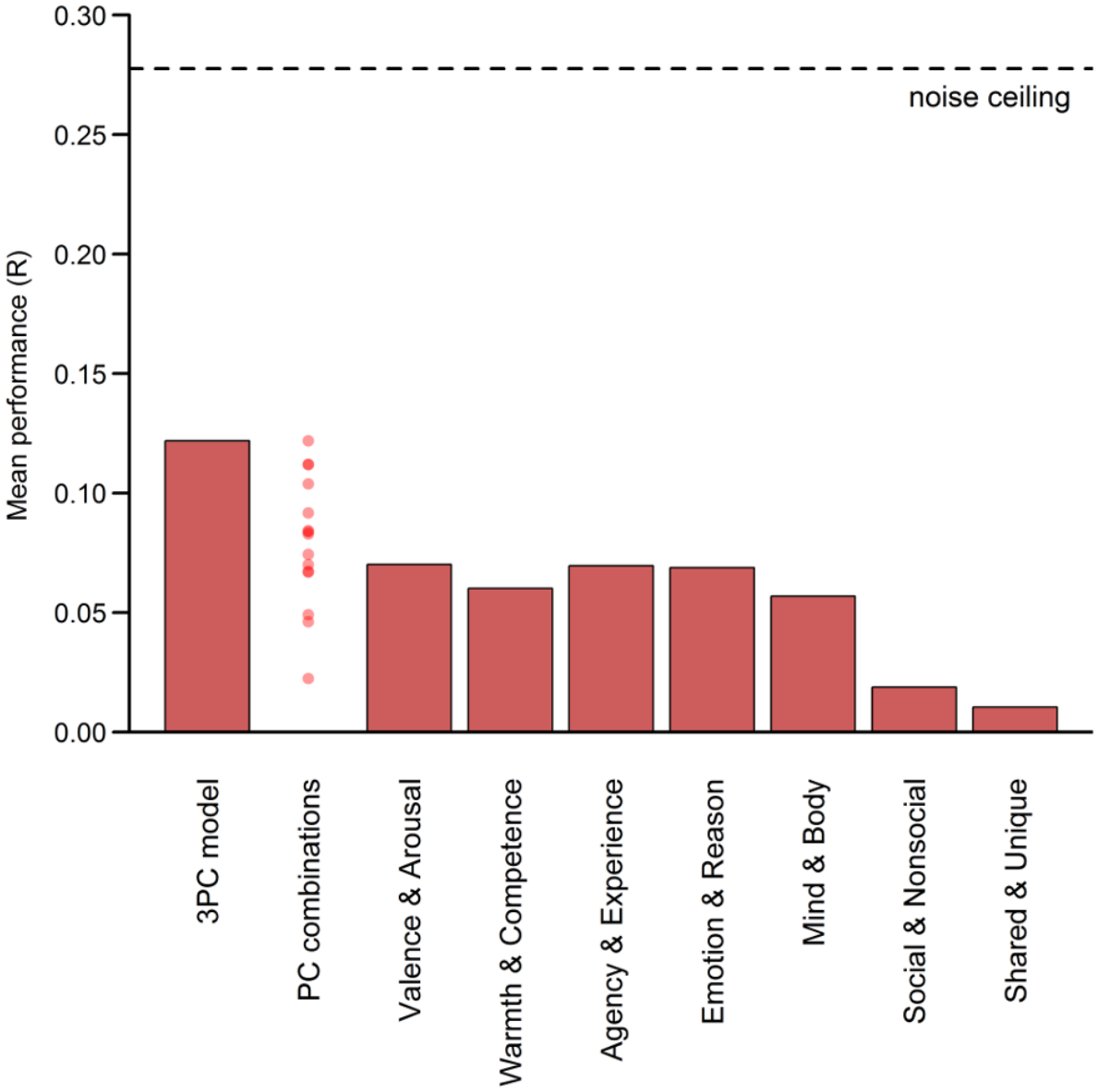


Figure 1.9 *Cross-validated model performance. Bars* indicate performance of a representation similarity analysis based on non-negative least-squares regression. Weights for dimensions within each theory were trained on data from 19 participants. This regression model was then tested by predicting the neural pattern similarity of the left out participant. Each participants was left out iteratively and results were averaged across all 20 training-testing combinations. Points in the "PC combinations" column indicate the performance of every possible combination of 1-4

Figure 1.9 (Continued). of the 4 PCs. The farthest left bar indicate the performance of the best model, consisting of the PCs rationality, social impact, and valence. The noise ceiling indicates the expected performance of an ideal model for mental state representation.

Neural patterns representing theoretical models. Though the primary purpose of this study was to discover the organization of mental state representation, we also tested whether the seven psychological theories from which we drew our PCA dimensions could predict neural representations of mental states. To do so, we repeated the wholebrain and network-level representational similarity analysis with the original psychological dimensions. Wholebrain analyses on each of the seven extant theoretical models revealed regions of the brain within which patterns of neural activity were predicted by each model (Figure 1.10 and Table 1.3). The valence and arousal model (Figure 1.10a) predicted patterns of activity in a number of regions including PCC, ACC, bilateral lateral temporoparietal cortex, left lateral and anterior temporal cortex, bilateral DLPFC, and both rostral and caudal portions of DMPFC. The warmth and competence model (Figure 1.10b) predicted patterns of activity in left TPJ, rostral and caudal DMPFC, bilateral ATL, bilateral VLPFC, and bilateral DLPFC. Agency and experience (Figure 1.10c) and emotion reason (Figure 1.10d) produced very similar results, an unsurprising outcome given the degree of correlation between these models. These models both predicted patterns of activity in VMPFC, rostral DMPFC, bilateral ATL, bilateral VLPFC and DLPFC, and portions of lateral temporal cortex. The mind and body dimensions (Figure 1.10e) predicted patterns in a proximal but distinct set of regions to those discussed above, including ACC, PCC, TPJ and portions of lateral prefrontal cortex. Sociality (Figure 1.10f) and human uniqueness (Figure

1.10g) models both predicted much less extensive clusters of activity, with both appearing in the precuneus and uniqueness also appearing in a posterior portion of the parahippocampal gyrus.

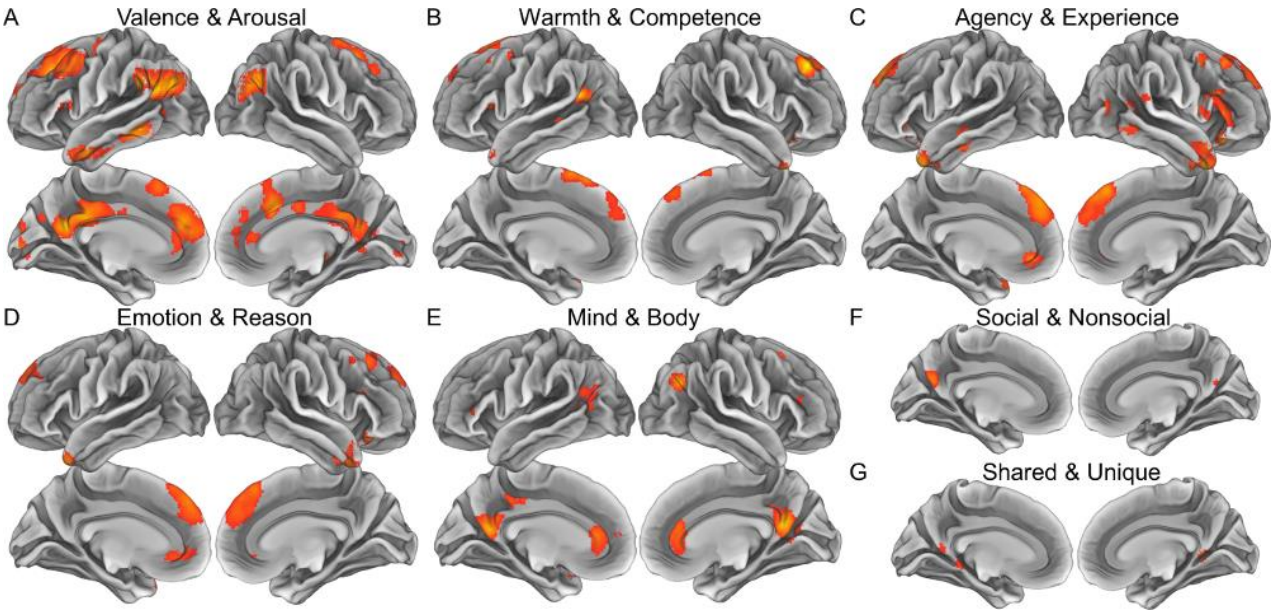


Figure 1.10 *Searchlight results for existing theoretical models.* The similarity of patterns within the yellow/orange regions can be explained by their proximity to each other on the dimensions of the corresponding social cognitive models ($p < .05$ corrected). Searchlight analyses were conducted on each participant and combined through one-sample random-effects *t*-tests.

Finally, we tested the degree to which of each of the seven theoretical models predicted patterns of neural activity in a distributed manner. At the network level, the predictions of 5 of 7 theoretical models were significantly correlated with neural similarity (Figure 1.5c): valence and arousal ($r = .19$, 95% bootstrap CI [.10, .23]), warmth and competence ($r = .16$, 95% bootstrap CI [.07, .21]), agency and experience ($r = .19$, 95% bootstrap CI [.09, .22]), emotion and reason ($r = .17$, 95% bootstrap CI [.06, .22]), and mind and body ($r = .18$, 95% bootstrap CI [.09, .22]), – all with statistically indistinguishable effect sizes. Two theoretical models did not predict network level patterns: social vs. nonsocial ($r = .04$, 95% bootstrap CI [-.03, .09]) and shared vs.

unique (r = .03, 95% bootstrap CI [-.003, .06]). These results were largely unchanged in the Spearman correlation variant: valence and arousal (r = .19), warmth and competence (r = .16), agency and experience (r = .19), emotion and reason (r = .17), mind and body (r = .18), social and nonsocial (r = .04), and shared and unique (r = .03). They were also similar in the alternative (NeuroSynth) feature analysis variant: valence and arousal (r = .16), warmth and competence (r = .17), agency and experience (r = .23), emotion and reason (r = .21), mind and body (r = .16), social and nonsocial (r = .04), and shared and unique (r = .02).

Table 1.3 *Regions containing patterns consistent with existing psychological theories*

Dimensions/Anatomical Label	x	y	z	Max T	Volume
Agency & Experience					
Anterior temporal lobe	-42	19	-40	6.17	462
Anterior temporal lobe/dorsomedial prefrontal cortex/dorsolateral prefrontal cortex	46	13	-36	5.81	5826
Superior temporal sulcus	-54	-19	-18	4.98	340
Lateral orbitofrontal cortex	-34	37	-10	5.21	281
Ventromedial prefrontal cortex	-6	41	-6	5.03	303
Temporoparietal junction/posterior superior temporal sulcus	58	-47	20	4.74	662
Valence & Arousal					
Anterior temporal lobe	-56	1	-30	5.33	531
Middle temporal gyrus	-58	-45	-10	5.37	608
Hippocampus	-34	-13	-12	4.42	78
Thalamus	12	-23	-4	5.29	120
Cuneus	-4	-89	0	5.49	282
Medial parietal lobe	-10	-55	20	6.5	2746
Dorsolateral/dorsomedial prefrontal cortex	8	21	34	6.82	7608
Temporoparietal junction	-46	-51	26	7.02	2311
Ventrolateral prefrontal cortex	-56	17	10	4.18	113
Temporoparietal junction	52	-69	30	6.01	960
Mid-cingulate gyrus	6	-1	38	4.41	137
Emotion & Reason					
Anterior temporal lobe	-42	19	-38	5.51	170
Anterior temporal lobe/orbitofrontal cortex	16	35	-8	5.95	951

Table 1.3 (Continued)

Ventromedial prefrontal cortex	-8	41	-4	4.18	310
Lateral prefrontal cortex	48	21	22	4.03	77
Dorsomedial prefrontal cortex	0	51	30	4.9	3011
Mind & Body					
Anterior insula	-40	13	-22	4.25	78
Medial prefrontal/Anterior cingulate cortex	0	35	12	4.95	916
Ventrolateral prefrontal cortex	-46	33	4	4.38	106
Temporoparietal junction	-42	-61	8	4.48	336
Medial parietal lobe	4	-53	20	6.23	1753
Dorsolateral prefrontal cortex	38	33	10	4.2	185
Temporoparietal junction	46	-67	32	5.68	316
Dorsolateral prefrontal cortex	22	15	52	4.39	194
Shared &. Unique					
Parahippocampal gyrus	-16	-39	-2	4.56	86
Posterior cingulate cortex	12	-49	6	5.63	260
Social & Nonsocial					
Precuneus	-4	-65	26	4.92	186
Warmth & Competence					
Anterior temporal lobe	42	19	-42	5.61	101
Anterior temporal lobe/ventrolateral prefrontal cortex	-50	19	12	4.3	374
Superior temporal gyrus	46	21	-12	4.05	94
Middle temporal gyrus	-48	-41	-4	4.31	135
Temporoparietal junction	-46	-57	20	6.02	371
Dorsomedial prefrontal cortex/dorsolateral prefrontal cortex	18	37	46	6.6	2638

Coordinates refer to the Montreal Neurological Institute stereotaxic space.

Discussion

The current study used fMRI and representational similarity analysis to explore the dimensions that organize our representations of other people's internal mental states. We used dimensions from the existing psychological literature on mental states as a springboard for generating four non-redundant, easily interpretable, dimensions, and tested which dimensions organize patterns of neural activity elicited by considering others' mental states. Results indicated that neural

activity patterns within the network of regions sensitive to others' mental states are attuned to three dimensions: rationality, social impact, and valence. These dimensions account for nearly half of the variation in the neural representation of mental states, constituting the most comprehensive theory to date regarding how we understand others' minds.

What significance do these three dimensions hold? One of these dimensions, termed "rationality," has arisen across disparate philosophical and psychological traditions. Here it derives from theories in the domain of social cognition, including primarily experience and agency (Gray et al., 2007), warmth vs. competence (Cuddy et al., 2008; Fiske et al., 2002), and emotion vs. reason, an idea extending back at least as far as Plato. This dimension may also closely mirror theories outside the social domain, such as active vs. passive (Osgood, Suci, & Tannenbaum, 1957), System I vs. System II (Kahneman, 2003), and reflective vs. reflexive (Heckhausen & Gollwitzer, 1987). The ubiquity of this distinction hints that it may reflect a deep principle of cognition. The results of the present study align with previous MVPA work (Corradi-Dell'Acqua, Hofstetter, & Vuilleumier, 2014) in suggesting that the brain spontaneously attunes to others' rationality. Knowing whether is experiencing a rational state or not may be particularly useful for certain social calculations. For example, it seems plausible that rationality assessments may help guide our decisions about whether people are responsible for their actions. This in turn would shape the degree to which we take those actions into account during impression formation. These functions have been repeatedly associated with DMPFC, one of the regions implicated in representing rationality (Mende-Siedlecki, Cai, & Todorov, 2013; J. P. Mitchell, Cloutier, Banaji, & Macrae, 2006; J. P. Mitchell, Macrae, et al., 2005; Schiller, Freeman, Mitchell, Uleman, & Phelps, 2009).

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

The third dimension to emerge from this study, "valence," captures the difference between positive and negative mental states. This concept has long been implicated in social and affective processing (Russell, 1980). As such, it may come as no surprise that valence plays an important role in the organization of mental state representations. Of note, however, is that we find a unique spatial distribution associated with this dimension. Previous work has associated the processing of positive vs. negative stimuli with specific neural networks, including the mesolimbic dopamine system (Sabatinelli, Bradley, Lang, Costa, & Versace, 2007), as well as other limbic structures, such as the amygdala (Garavan, Pendergrass, Ross, Stein, & Risinger, 2001). Supplementary univariate analyses do show that the VMPFC, a region involved reward

and value more generally, tracks the positivity of mental states. However, our MVPA results did not identify these regions, but instead implicated left-lateralized cortical regions in lateral prefrontal cortex and the angular gyrus. One possible explanation is that language supports the processing of mental state valence, but not other types of valence, a hypothesis here only preliminarily supported by the lateralization and the proximity of the valence regions to known language areas.

Together the three significant dimensions described above explain a nearly half of the reliable variance in the neural representation of mental states. The social impact dimension alone predicts as more variance than the original theoretical models from which we derived our new dimensions; the combination of the three significant PCA-derived dimensions explains approximately twice the variance of the circumplex model, the most successful of the original theories. At the same time, given their significance to psychological theory, it is both reassuring and unsurprising that five of the seven original theories significantly predict neural pattern similarity. Notably, even theories that were originally geared towards explaining traits or groups, such as the stereotype content model, demonstrate their efficacy in the mental state domain. This raises the interesting possibility that the same dimensions organize neural activity about different types of social construct.

In addition to informing us about the psychological question of interest – the organization of mental states – the current results also hint at the neural encoding scheme within the social brain network. By assessing the representation of mental states at two different levels of spatial organization – local activity patterns within spherical searchlights and broader activity patterns across the social brain network – the current study is well placed to bear on this issue. The results of the present study support the hypothesis that information is encoded by patterns of activity

within localized brain regions, rather than *across different regions*. If local patterns did not encode social information but coarse patterns across the network did, the searchlight analysis would fail to produce results. Instead, we observe reliable encoding of mental state information in local patterns across the social brain, and explanatory power at the network level appears roughly in proportion to the cortical extent of their local encoding. As such, the current results provide no evidence that others' mental states are represented by inter-regional activity differences above and beyond the information already contained in local patterns. Interestingly, we find that two regions, the dMPFC and TPJ, each underlie multiple dimensions. Previous work has already identified heavily implicated these regions in mentalizing. The convergence of multiple dimensions on these nodes may help to explain their prominence in this domain.

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

We can also ask how these dimensions might apply across social cognition more generally. The current study used only lexical stimuli, and tested these on only English-speaking

adults. Do these dimensions apply to social cognition in other cultures? Do infants, or other primates demonstrate any of the building blocks of these dimensions? Do these same dimensions apply when mentalizing about non-linguistic content? Previous work on cross-modal emotion representation indeed suggests that visual and verbal emotional stimuli may be processed similarly (Peelen, Atkinson, & Vuilleumier, 2010; Skerry & Saxe, 2014) though the full model has yet to be tested. We hope that the current data will provide a solid foundation for future research in these domains. It is also worth considering precisely which processes the imaging task taps. The task relies heavily on conceptual representations of mental states, and it is not entirely clear how strongly these concepts might guide other forms of mentalizing.

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

The present study derived four potential dimensions of mental state representation – rationality, social impact, human mind, and valence – from the existing psychological literature. We discovered that three of these dimensions – rationality, social impact, and valence – predicted patterns of neural activity elicited across the social brain network by consideration of others' mental states. By discovering which dimensions the brain spontaneously uses to organize the domain of mental states, we have forged a deeper understanding of both human social

cognition and its relationship to our own internal mental experience. These findings both inform long-standing debates within social psychology about theory of mind, and can be used to generate novel predictions about how the brain supports our ability to mentalize.

Paper 2: Consistent neural activity patterns represent personally familiar people

Thornton, M. A., & Mitchell, J. P. (accepted). Consistent neural activity patterns represent personally familiar people. *Journal of Cognitive Neuroscience.*

The individual person is a fundamental unit of social life. To navigate our own social lives, we must have a detailed understanding of the individuals who populate it, including an ability to anticipate their idiosyncratic thoughts, feelings, and actions. As such, the mind and brain must represent knowledge of not just general human psychology – a lay theory of mind – but also of the particular psychologies of people who are important to us. Such person-specific knowledge could guide both direct interpersonal interactions as well as offline simulations of others' thoughts and actions. The present investigation aimed to characterize the neural encoding of our knowledge of personally familiar others. In addition to investigating which brain regions support person-specific representations, we examined how such representations contribute to the process of simulating other minds. In particular, we identified brain regions that are likely candidates for integrating person information with context information to form complete simulations of social events. We also tested the hypothesis that instating a 'typical' person-specific pattern serves as a meta-cognitive clue to the predictive validity of such simulations. Finally, we examined several theories of person perception to determine which, if any, provided an appropriate taxonomy for our neural representations of personally familiar others.

Several previous studies have investigated the neural representation of personally familiar individuals. Regions including medial prefrontal cortex, medial parietal cortex, the temporoparietal junction, and the superior temporal sulcus have been consistently implicated in social thought, forming the so-called "social brain" network (for review, see J. P. Mitchell, 2008; Van Overwalle & Baetens, 2009). It is possible that many or all of these regions support person-specific representations, a position supported by studies examining the neural correlates of

(person) familiarity (Cloutier, Kelley, & Heatherton, 2011; Gobbini, Leibenluft, Santiago, & Haxby, 2004). Evidence from an fMRI adaptation (i.e., repetition suppression) study also suggests that a broad array of brain regions may contain information specific to thinking about particular targets (Szpunar, Jacques, Robbins, Wig, & Schacter, 2014). However, both adaptation and activation-based fMRI studies have also suggested a more specific role for ventral medial prefrontal cortex in representations of specific familiar others (Heleven & Van Overwalle, 2016; Welborn & Lieberman, 2015).

A reasonable question one might ask on the basis of such studies is to what extent the representation of other people is distributed across the social brain versus supported by a single, dedicated module? This debate over distributed versus modular functioning has long occupied cognitive neuroscientists generally (de Beeck, Haushofer, & Kanwisher, 2008). Understanding the functional brain architecture supporting representation of specific persons might help address questions of interest to psychologists. For example, if person-specific information appeared to be distributed throughout the social brain network, as previous studies have suggested, it would suggest that detailed knowledge about other people permeates our social cognitive machinery. In other words, it would hint that even social judgments that do not necessarily require any background knowledge may make use of person-specific information when it is available.

Identifying representations of individual people presents a challenge because traditional techniques for analyzing functional magnetic resonance imaging (fMRI) data are ill-suited to the task. The social brain network activates very differently in response to social and non-social stimuli, but imagining a set of familiar people would likely elicit comparatively similar levels of activity for each individual. Thus traditional univariate approaches, which rely on differences in average activity levels to distinguish between stimuli, may prove unable to differentiate target

people, even in brain regions that *do* contain person-specific representations. Fortunately, a nascent set of tools for exploring fMRI data may now make it possible to identify where these representations reside. These techniques, known collectively as multivoxel pattern analysis (MVPA), focus on fine-grained patterns of neural activity within brain regions, or coarse-grained activity patterns across multiple regions, rather than on gross average activity levels within individual regions (Haxby et al., 2001). As a result, MPVA can achieve greater sensitivity than traditional univariate statistical analyses and can detect multidimensional neural codes that may exist even in the absence of large-scale differences in average neural activity (Davis et al., 2014).

Here we apply the form of MVPA known as representational similarity analysis (RSA) to identify brain regions that demonstrate different patterns of neural activity when perceivers imagine different people (Kriegeskorte, Mur, & Bandettini, 2008). We entered into the investigation agnostic to the spatial scale at which person-specific information might be encoded, and thus we applied RSA in search of both fine-grained intra-regional activity patterns – via the use of searchlight mapping (Kriegeskorte et al., 2006) – and coarse-grained inter-regional patterns distributed across the social brain network as a whole.

An initial MVPA-based foray into the representations of individual people has already yielded promising results (Hassabis et al., 2014). In that study, patterns elicited by thinking about four fictional people in an episodic simulation task were reliably classified in two adjacent portions of dorsal medial prefrontal cortex. The biographies of the fictional targets were written to create strong impressions of certain traits (agreeableness and extraversion). Additional brain regions were capable of these decoding individual personality traits, but only the medial prefrontal regions could distinguish between all four target people. It remains unclear whether the classification was being driven by the sort of rich, detailed knowledge that characterizes our

representations of personally familiar others, or simply by the exaggerated social differences created by this experiment's central manipulation. To resolve this uncertainty, the present study used a broad set of personally familiar targets.

The present study was designed to cast light on more than just which brain regions support representations of specific people. When participants in our experiment imagined people they knew, they imagined them within a set of different contexts. This allowed us to assess which brain regions support context-specific patterns of neural activity, and where representations of person and context identity overlap. The question of how knowledge of dispositions and situations are integrated when making attributions has been of long standing interest to social psychology (Gilbert & Malone, 1995; Jones & Harris, 1967). Recent neuroscientific investigations have suggested that increased activity in the social brain network predicts dispositional attributions, but have been inconsistent regarding the positive predictors of situational distributions (Kestemont, Vandekerckhove, Ma, Van Hoeck, & Van Overwalle, 2013; Moran, Jolly, & Mitchell, 2014). By focusing on fine-grained patterns rather than on raw activity levels, the present study aimed to provide a clearer view of which region(s) are potential loci for the integration of situational and dispositional information. While our design does not permit a detailed assay of the nature of such integration, localizing regions that support both types of information serves as a crucial first step in this direction.

By examining mental simulations of the same set of people over several iterations, we also aimed to shed light on the psychological process of imagining other people. In particular, we analyzed the patterns of activity during each simulation to reveal how perceivers arrive at meta-cognitive judgments of confidence in their imagination. Over the course of the study, participants were presumably able to integrate their knowledge of target people into some simulations better

than others. For example, imagining an elderly relative having dinner with you is not particularly outlandish, whereas picturing that same person frolicking at the beach may prove rather more difficult. We hypothesized that the degree to which one can easily incorporate knowledge of a person into a simulation could signal how valid that simulation might be: that is, how likely to accurately reflect the real actions of a given person in a given situation. In the present paradigm, we measured the degree to which the pattern elicited on each trial resembles the patterns across all other trials featuring the same target person. This pattern typicality measure served as an indirect index of the incorporation of person knowledge into each simulation. We expected pattern typicality to be positively associated with trial-wise ratings of vividness and accuracy.

Finally, we planned to examine several theories which might explain the differences in person-specific neural representations. Determining how people naturally divide up the social world has produced many valuable theories in social psychology. Although these theories can each explain a variety of social behaviors, it is unclear which taxonomies characterize the information that perceivers spontaneously draw upon to simulate others' minds. To address this question, we examined a wide range of theories as candidate explanations for the similarities between person-specific activity patterns. These candidates included five major conceptions of person perception, self-reported and implicit measures of holistic similarity, and in-scanner ratings of accuracy and vividness. The extant social theories were the following: categorization of basic social groups (age, race, sex); the Five Factor Model of personality (Goldberg, 1990; McCrae & Costa, 1987), consisting of openness, conscientiousness, extraversion, agreeableness, and neuroticism; the dimensions of warmth and competence, which characterize the Stereotype Content Model (Fiske et al., 2002); egocentric factors (similarity, familiarity, and liking); and

dyadic relational model (sharing, trading, or ordering) between target and participant (Haslam, 2004; Haslam & Fiske, 1999).

Each of these different characterizations of the target people makes different predictions about which targets should elicit more similar or more different neural activity patterns. For example, two target people might belong to the same basic social groups, but have very different personalities on the Big 5 factors. The social groups model would thus predict similar patterns of neural activity for these two targets, while the Five Factor Model would predict relatively dissimilar patterns. Using representational similarity analysis (Kriegeskorte, Mur, & Bandettini, 2008), we were able to compare these theory-based predictions about inter-target similarity with the actual similarity between patterns of neural activity. Thus we could test hundreds of predictions simultaneously to determine which theory best accounts for the observed neural data. Directly comparing the theories to each other would require much more statistical power, and so in the present study we instead focused on examining whether there was evidence for each theory's influence on pattern similarity at all. We thus attempted to select as broad a range of theories as possible rather than a set of highly similar, competing theories.

Methods

Raw behavioral data, processed imaging data, and custom experiment presentation and statistical analysis code for this study are freely available on the Open Science Framework (http://osf.io/gxhkr/). Raw imaging data have been deposited at the Harvard University Dataverse (http://dx.doi.org/10.7910/DVN/ZQQABJ). Shared data have been stripped of identifying features for the privacy of the participants.

Participants. A power analysis was conducted via Monte-Carlo simulation to assess the number of participants and trials needed with the design specified below. This analysis targeted

the final planned representational similarity analysis, in which inter-target pattern similarity was to be modelled in terms of existing theories of person perception. We targeted this analysis rather than the primary trial-wise representational similarity analysis to avoid the additional uncertainly of estimating context and run effect sizes at the trial level. Simulated patterns of neural activity were generated to embody an anticipated effect size of $r = .15$ (correlation between model and neural data) which we judged to be reasonable based on previous research (Kriegeskorte, Mur, & Bandettini, 2008; Kriegeskorte, Mur, Ruff, et al., 2008). This was achieved by creating a noise free "model" pattern $\sim N(0, 1)$ and adding noise to create a simulated neural pattern matrix with the appropriate relationship to the model. A simulation-based search across noise parameters determined that a $\sim N(0, 2.4)$ distribution of noise yielded the appropriate effect size.

Additional, independently generated noise $\sim N(0, 3.4)$ was then added to these simulated neural patterns to produce patterns of activity for each simulated experimental trial. The trial-wise noise parameter was calculated by adjusting the same parameter used in a similar power analysis for a previous study (Tamir, Thornton, Contreras, & Mitchell, 2016) to account for the difference in trial duration (12.5s modelled in this study vs. 4.25 in the previous study). Each simulated pattern set consisted of 20 (target people) by 200 (voxel) elements. Since we could not anticipate the size of activations in advance, voxel count was set to near the average searchlight size. Patterns were then subjected to representational similarity analysis as described below under imaging procedures, producing 20 x 20 similarity matrices for each simulated participant that could be compared to the correlation matrix of the original "model" pattern. To simplify computation, the general linear model phase of this analysis was reduced to averaging the patterns within each participant across trials. This process was repeated 100 times with simulated participant numbers ranging from 2 to 25 and trial numbers ranging from 2 to 10 per target.

These simulations indicated that 25 subjects with 6 trials per person should be adequate to ensure 95% statistical power at a threshold of $\alpha = .001$.

Participants ($N = 25$) were recruited from the Harvard University Psychology Study Pool. Two were subsequently excluded: one due to a neurological anomaly detected during scanning, the other due to excessive head motion. All remaining participants (15 female; age range = 18–27, mean age = 21.1) were right-handed, neurologically normal, fluent in English, and had normal or corrected-to-normal vision. Participants provided informed consent in a manner approved by the Committee on the Use of Human Subjects in Research at Harvard University.

Stimuli and behavioral procedure.

Pre-testing. We selected a set of 20 target stimuli to maximize observable differences in the imaging experiment. Participants first provided the names of 40 personally familiar adults. To choose a highly diverse set of targets, and to provide measures for representational similarity analysis, we asked participants to rate these people on 16 social dimensions. These included Big 5 personality traits (Goldberg, 1990; McCrae & Costa, 1987), warmth and competence (Fiske et al., 2002), the degree to which their relationship with the person was predicated on communal sharing, equity matching, and authority (Haslam, 2004; Haslam & Fiske, 1999), and similarity, familiarity and liking. Participants also indicated the race, sex, and age of each target. Single item ratings on 7-point Likert scales were used for all dimensions other than the demographics. Multi-item scales would have been preferable psychometrically, but inordinately time-consuming and exhausting for participants. We provided definitions of the dimensions to ensure that participants used the scales consistently. The dimensions were presented in a unique random order for each participant, and the order of the target people was randomized for each dimension.

These ratings were used to select a subset of 20 target people who were diverse across all 16 dimensions. The selection process proceeded as follows: first the target most different (in Euclidean distance) from centroid of the group across the 16 measured dimensions was selected; on each subsequent iteration of the algorithm, the target person who would maximize the theory-weighted average standard deviation of the selected set would be added to it. The algorithm terminated once 20 targets had been selected. Although this procedure was not guaranteed to produce an optimal selection, it was very rapid to compute – a tradeoff we deemed worthwhile given that it had to be performed while participants waited. This approach succeeded in producing an increase in inter-target variance in 20 of 23 participants, and yielded only minor decreases in variance in the other three.

Participants also provided holistic similarity measures in a separate behavioral task. On each of 380 trials, participants compared two target people to one reference target and indicated which of the two was more similar to the reference. The trial number allowed for two presentations of each unique pair of target people (in the non-reference positions). Participants' choices were used to derive a self-reported explicit holistic similarity matrix between the targets. This similarity matrix was generated by summing the times a particular pair of reference and selected comparison targets were judged to be the more similar of the two possible pairs, and dividing this sum by the number of possible trials on which this judgment might have occurred. Reaction times were used to generate an equivalent implicit holistic similarity matrix, with longer reaction times indicating greater similarity between the two choice targets.

Experimental paradigm. Just prior to scanning, participants were asked to imagine each of six common situations: taking a long road trip, shopping for groceries, going to a fair, going to the beach, waiting for the doctor, having dinner at a restaurant. They were instructed to imagine

these situations as vividly as possible from their own perspective. During fMRI scanning, participants simulated each of the 20 target people in each of these contexts (Hassabis et al., 2014). Each trial began with a prompt (2.5 s) indicating which context the participant should simulate (e.g., "eating at a restaurant"). Next, the name of one of the 20 targets appeared (2.5 s), after which the screen went blank and participants had 10 s to simulate that target in the specified context. Participants were instructed not to imagine new situations on each trial, but instead to place each target person in the same previously imagined context.

After the simulation period, participants rated the vividness of their simulation on that trial (2.5 s) and how accurate they felt their simulation was with respect to what the target person would actually do in that context (2.5 s). These ratings were averaged to form a "confidence" composite. The six contexts were fully crossed with the 20 target people for a total of 120 trials divided across 6 runs of 400 s each. Note that each combination of target person and context occurred only once, making the task a unique-trial design. Each target person was shown only once within each run, but the position of a given target person within a run was randomly determined (independently for each participant). The order of contexts was independently randomized for each target person within each participant. All experimental tasks were presented using Python 2.7 and the PsychoPy package (Peirce, 2007).

Imaging procedure. Functional gradient-echo echo-planar images were obtained from the wholebrain (43 interleaved slices of 2.5mm thickness parallel to the AC-PC line, TR = 2500 ms, TE = 30 ms, flip angle = 90°, in-plane resolution = 2.51 x 2.51 mm, field of view = 216 mm^2, matrix size = 86 x 86 voxels, 160 measurements per run) using a parallel imaging protocol and prospective acquisition correction (PACE). A high resolution T1-weighted structural scan (multi-echo MPRAGE, 1.195 mm isometric voxels, matrix size 192 x 192 voxels, 144 sagittal

slices) was also acquired from each participant. Imaging data were acquired at the Harvard University Center for Brain Science with a 3 Tesla Siemens Tim Trio scanner (Siemens, Erlangen, Germany) using a 32 channel head coil. Functional images were preprocessed and analyzed using SPM8 (Wellcome Department of Cognitive Neurology, London, UK) with the SPM8w extension (https://github.com/ddwagner/SPM8w) and in-house scripts in MATLAB and R. Data were first spatially realigned via rigid body transformation to correct for head motion and then normalized to a standard anatomical space (2 mm isotropic voxels) based on the ICBM 152 brain template (Montreal Neurological Institute). The general linear model (GLM) was used to analyze each participant's data in preparation for MVPA. Two separate GLMs were run, to allow for analysis at different levels: with respect to individual trials and at the level of target people.

In the first of these GLMs, each trial in the experiment was modelled separately (120 conditions of interest) as a boxcar regressor starting at the beginning of the target person presentation period and ending at the completion of the imagination period (12.5 s). Note that modelling only this period of each trial allowed for relatively long (7.5s) inter-trial intervals, minimizing multicollinearity in the design matrix. The boxcars regressors were convolved with a canonical hemodynamic response function and entered into the GLM. The model also included additional covariates of no interest: run trends and means, six motion realignment parameters, and outlier time points. The second GLM included only 20 conditions of interest, corresponding to the 20 personally familiar target people. Thus each boxcar regressor modelled 6 trials, corresponding to imagining each target person in each of the 6 contexts presented. Otherwise this GLM proceeded identically to the first, with the exception of adding temporal and dispersion derivatives of the conditions of interest to the model.

Searchlight analyses. We extracted local patterns of neural activity from throughout the brain using an approximately spherical searchlight with four voxel radius (~9 mm). In order to ensure that the edges of the brain were included, searchlights with up to half of their voxels implicitly masked (i.e. outside the brain) were analyzed. Thus searchlight size varied between 128 and 257 voxels. For each voxel in the brain, patterns of activity (*t*-value maps from the first GLM) were extracted from the surrounding searchlight area for each condition of interest (trial in the experiment). We estimated the similarity between activity patterns by calculating the Pearson correlation between the values in each pair of patterns. These estimates were then rank transformed and regressed onto two binary variables encoding our hypotheses. Rank transformation is a recommended procedure for representational similarity analysis, as it helps mitigate large scale differences and violations of distributional assumptions that may otherwise compromise results (Kriegeskorte, Mur, & Bandettini, 2008). The person-representation variable predicted high similarity between trials with the same target person and low similarity between trials with different targets. Analogously, the context-representation variable predicted high similarity between trials with the same context and low similarity between trials with different contexts. We also included a similarly-constructed run nuisance variables to control for pattern similarity between trials in the same run.

This regression analysis produced whole-brain unstandardized beta maps for each participant. These maps were subjected to smoothing with a Gaussian kernel (4 mm FWHM) to maximize inter-subject alignment and were then entered into a random effect analysis across participants using one-sample *t*-tests. This analysis was corrected for multiple comparisons using a MATLAB implementation (https://github.com/markallenthornton/MatlabTFCE) of maximal statistic permutation testing with threshold free cluster enhancement (TFCE; Smith & Nichols,

2009). The family-wise error rate was strictly controlled across three random effects t-tests from the two searchlight analyses (the person- and context-representation mappings described above, and the confidence analysis described below) by setting the permutation-corrected critical p-value within each map to a Bonferroni adjusted $\alpha = .0167$. Results were rendered on the cortical surface using Connectome Workbench (Marcus et al., 2011).

Person-pattern typicality was estimated for each trial by averaging the similarity values between the trial in question and the other trials involving the same target. High typicality for a trial thus meant that the pattern for this trial was highly correlated with the patterns for the other trials on which the same target was presented. Each measure in the pattern typicality vector was the mean of five correlations (since each person was imagined in six contexts total). For each searchlight region, the pattern typicality vector was entered into a multiple regression as the dependent variable to be predicted by the confidence composite ratings. Additional nuisance regressors for target person were also included to control for the possibility that some targets might elicit more reliable patterns than others. As with the primary searchlight, this analysis produced wholebrain unstandardized beta maps for each participant. These maps were smoothed with a 4 mm FWHM Gaussian and then entered into random effects *t*-tests, with multiple comparisons corrected as described above.

Feature-selected representation similarity analysis. As reported in detail below, the person-representation searchlight analysis revealed an extended set of regions containing fine-grained person-specific activity patterns. The statistically significant regions in this analysis were used as a mask to select voxels for further analysis. Note that this selection process, despite drawing on dependent data, does not constitute statistically biased "double-dipping" for the purposes of the analyses we apply. This is because the independent variables that will be used to

model the feature-selected patterns (i.e., participant ratings on the 16 social dimensions) were not themselves involved in voxel selection. While on average this procedure will increase the observed size of real effects by disattenuating correlations, it will not systematically lead to the generation of spurious relationships between the models and the data. To avoid circularity, we do not test any models using the full deconvolution (trial-wise) GLM within this mask.

In addition to this feature-selection approach – which should produce minimally reliability-attenuated estimates – we also repeated the same set of analyses using an independently defined mask from previous research (Tamir et al., 2016). The voxels in this mask were chosen by univariate differentiation of mental state concepts. Using this independent mask justifies stronger claims about the localization of representations within the regions sensitive to mentalizing *per se*, and allows us to test for the presence of person-specific patterns at the trial level without circularity. This alternative feature selection approach also allows us to demonstrate the unbiased nature of our reliability-based feature selection.

For the 15,849 retained voxels in the reliability selection (Figure 1a and Table 1), and the 25,215 voxels in the independent mask (Tamir et al., 2016), patterns of neural activity (unstandardized beta maps) were extracted from the second GLM: for each participant, one pattern of activity across the feature-selected regions for each target person. Similarity estimates between these neural representations were again calculated by Pearson correlating the activity patterns. These estimates were then Spearman correlated with predictions of inter-target similarity made by models described above: the five theories of person perception, the two measures of holistic similarity, and the Euclidean space described by in-scanner ratings of vividness and accuracy, averaged by target person. Predictions of (dis)similarity between targets for each of the five theories of person perception were calculated by taking the Euclidean

distance between ratings of target people within the n-dimensional space described the dimensions of the theory. In the case of the basic social groups theory, since the distance between categories is not a calculable quantity, the dissimilarity matrix was binary, with values indicating whether a pair of targets belonged to the same (sex, age, or race) group or not.

This representational similarity analysis generated correlation coefficients for each of the 7 models within each participant. The statistical significance of these results were assessed through complementary direct and indirect testing procedures. One-sample t-tests across participants were conducted on the coefficients for each model. Fisher's r-to-z transformation was applied prior to the correlations prior to conducting these t-tests. Additionally, non-parametric percentile bootstrapping was used to obtain robust 95% confidence intervals around the mean correlation coefficient for each model. Both testing procedures were conducted separately for the two feature selection methods.

The independent feature selection method allowed us to perform an additional test of our primary hypothesis: that patterns of activity in the social brain network encode personally familiar people. Repeating this analysis in the feature-selected regions had two primary advantages: it allowed for the unbiased estimation of a single summary measure of effect size, and it allowed us to examine whether person-specific patterns persist across spatial scales in the social brain network. This analysis was achieved by repeating the multiple regression representational similarity analysis described above. However, in this case, the entire set of voxels in the independent mask was used to produce a single results neural similarity matrix for each participant. Results were combined across participants via random effects t-tests on the regression coefficients from each participant. To determine whether this analysis depended on the same fine-grained patterns as the searchlight analysis or instead relied on coarse-grained

patterns, it was repeated with unsmoothed and smoothed patterns of brain activity. The heavy degree of smoothing applied in the latter case (18 mm FHWM, equal to the diameter of the searchlight) served to ensure that any effects observed could be attributed to coarse-grained inter-regional patterns rather than fine-grained intra-regional patterns.

Results

Searchlight results. Consistent patterns of neural activity for familiar others were observed across wide areas of cortex (Figure 2.1a). These areas primarily overlapped with regions previously implicated in social cognition, including medial prefrontal cortex (dorsal and ventral), the temporoparietal junction and superior temporal sulcus bilaterally, and most robustly, medial parietal cortex including portions of both precuneus and posterior cingulate (Table 2.1). Person-specific patterns were also observed in right ventral lateral prefrontal cortex, bilateral dorsal prefrontal cortex, medial precentral gyrus, and a portion of the posterior insula.

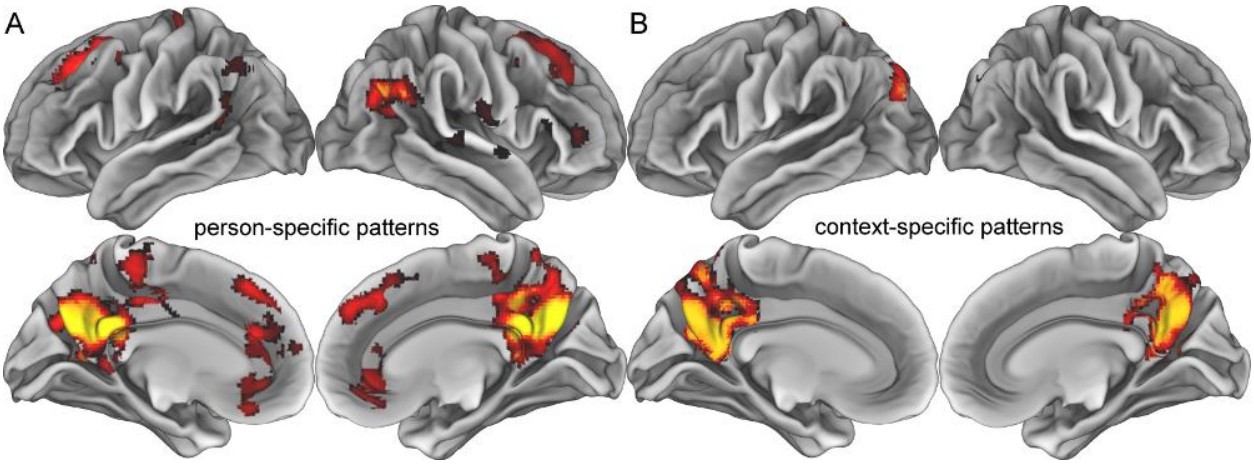


Figure 2.1 *Person- and context-specific patterns.* Consistent patterns of neural activity were elicited in the red/yellow regions during mental simulation for target people (A) or the context in which they were imagined (B). Results from searchlight mapping were combined across participants via t-test and corrected for multiple comparison via maximal statistic permutation testing with threshold free cluster enhancement.

Context-specific patterns were observed in a more circumscribed set of regions (Figure 2.1b). These included medial parietal cortex and bilateral superior occipital cortex (Table 2.1). Medial parietal cortex – precuneus in particular – has long been implicated in mental imagery (Cavanna & Trimble, 2006; Fletcher et al., 1995), which may account for its role in representing context. The occipital areas may overlap with or be adjacent to the transverse occipital sulcus, which has been implicated in scene perception (Bettencourt & Xu, 2013). This would be consistent with their implication in context individuation in the present study. Note that the only region of overlap between person- and context-specific patterns was in the medial parietal cortex.

Table 2.1 *Peak Voxel and Cluster Size for All Regions Obtained from the Searchlight Mappings.*

Map/Anatomical Label	x	y	z	Volume	Peak P
Person-specific patterns					
Medial parietal cortex/Right temporoparietal junction	-2	-45	22	8651	<.0001
Medial prefrontal cortex	-18	25	44	5256	.0045
Left temporoparietal junction	-44	-51	18	883	.0121
Right inferior frontal gyrus	40	39	4	651	.0115
Right superior temporal sulcus	48	-33	-6	391	.0130
Right Posterior Insula	30	-25	12	17	.0163
Context-specific patterns					
Medial parietal cortex /Left superior occipital cortex	-6	-53	14	5968	.0010
Right superior occipital cortex	32	-75	32	207	.0121
Confidence-pattern typicality correlation					
Precuneus	14	-65	38	551	.0053
Posterior/mid-cingulate	2	-21	34	372	.0120
Precuneus	22	-51	44	15	.0158

Coordinates refer to MNI stereotaxic space. Volume refers to voxel number. Peak P indicates TFCE-corrected p-value at peak.

A secondary searchlight analysis aimed to assess the relation between person-specific pattern typicality and by-trial ratings of confidence (accuracy and vividness). A positive relation between pattern typicality and confidence was detected in three adjacent regions within medial parietal cortex (Table 2.1). The most anterior of these regions was placed along the posterior

cingulate gyrus, while the other two areas were located primarily within the precuneus (Figure 2.2). Within these regions, the more typical a pattern was for a particular target the more confidence participants would express in their simulation on a given trial. This result suggests both neural and psychological interpretations. Neurologically, it suggests that medial parietal lobe plays a particularly important role in actively 'running' mental simulations, or at least making them consciously available. Psychologically, this result suggests that people may use the degree to which they can integrate idiosyncratic person knowledge into a simulation as an indicator of simulation validity.

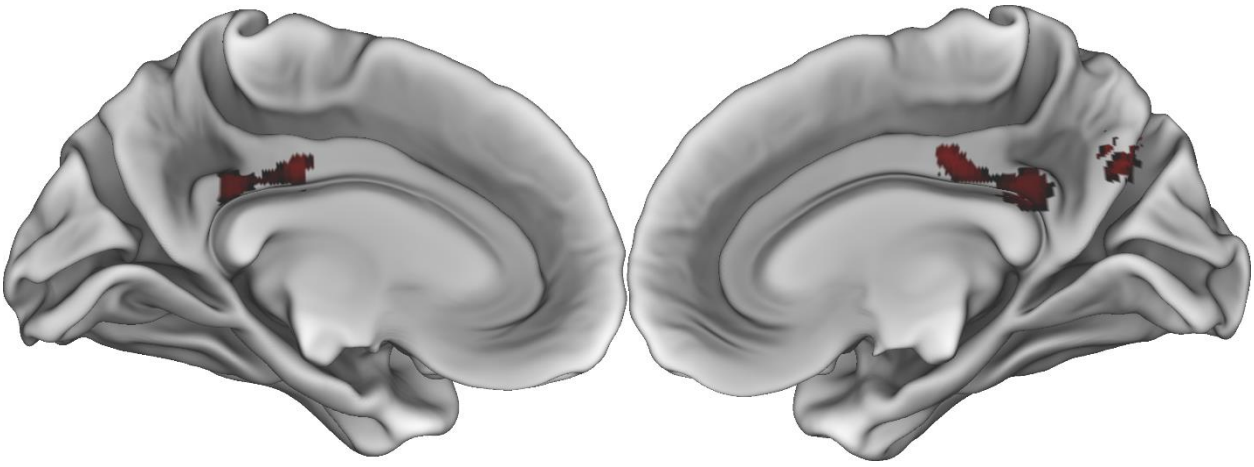


Figure 2.2 *Pattern-typicality predicts confidence.* The regions in red showed a positive relation between the instatement of typical person-specific patterns and by-trial ratings of simulation vividness and perceived accuracy. Results from searchlight mapping were combined across participants via t-test and corrected for multiple comparison via maximal statistic permutation testing with threshold free cluster enhancement.

Feature-selected representational similarity results. Person-specific activity patterns were extracted from the significant voxels in the corresponding searchlight analysis (Figure 1a) or using an independent mask for voxels sensitive to mental state representation (Tamir et al.,

2016). The similarity between person-specific patterns was modelled in terms of feature sets via representational similarity analysis. The results of this analysis provided consistent evidence that simulation confidence and holistic explicit similarity judgments predicted the neural representations of personally familiar others (Figure 2.3).

The average confidence (vividness and accuracy) of simulations involving each target was the most robust predictor of pattern similarity for both reliability-selected (mean r = .08, d = .66, p = .005) and independent (mean r = .09, d = .82, p = .0007) versions of the analysis. Although not contributing theoretical detail, self-reported behavioral ratings of holistic similarity also predicted pattern similarity in both the reliability selected regions (mean r = .03, d = .48, p = .03) and the independently selected regions (mean r = .03, d = .58, p = .01). Holistic implicit similarity was descriptively weakly positively associated with pattern similarity, but not at a significant level for either type of feature selection (ps > .1).

Results indicated weak evidence for the influence of several theories of person perception on neural pattern similarity (Figure 2.3). Egocentrism appeared to have a significant effect in the independent regions (mean r = .04, d = .50, p = .03), but was slightly less correlated with pattern similarity in the reliability selected areas and was not a statistical significant predictor therein(mean r = .03, d = .34, p = .12). Big 5 personality traits predicted pattern similarity at a marginally significant level under both the reliability-based (mean r = .04, d = .41, p = .06) and independent (mean r = .04, d = .36, p = .098) feature selection regimes. Relational models theory was a significant predictor of target pattern similarity in the reliability-based analysis (mean r = .05, d = .49, p = .03), but was only marginally significant for voxels within the independent mask (mean r = .05, d = .40, p = .07). Stereotype content was a marginally significant predictor, though only in the independent feature-selection analysis (mean r = .04, d = .38, p = .08) and not

the voxels chosen via reliability (mean $r = .02$, $d = .19$, $p = .36$). Similarity in terms of basic social groups (age, race, and sex) was slightly negatively correlated with pattern similarity, though the results were not significantly different from zero.

Despite some differences in categorical statistical significance due to near-threshold p-values, the representationally similarity analysis results were generally highly similar across the two feature selection methods. Direct comparison between the theories was not undertaken, as the study was not designed with sufficient power for that purpose. Descriptive comparison of the models should also be approached cautiously, as differences in effect size may result from differences in reliability across models or – in the case of the accuracy and vividness judgments, greater psychological proximity to the simulations.

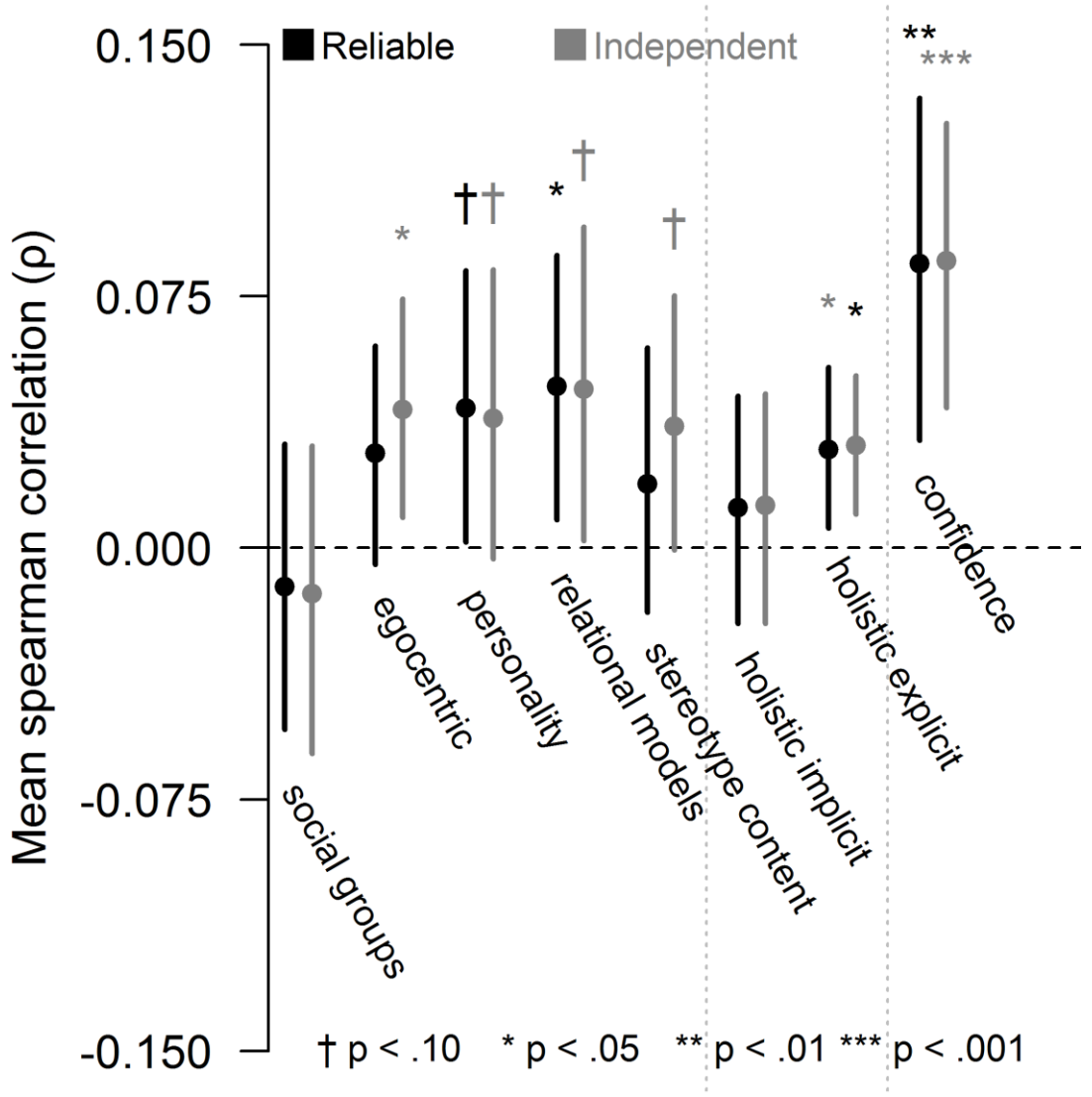


Figure 2.3 *Representational similarity analysis.* Person-specific patterns of neural activity across the feature-selected voxels were modelled in terms of a number of possible predictors. Two different forms of feature-selection were used: reliability-based selection was performed by using voxels which significantly encoded target people in the earlier searchlight results, and an independent mask of voxels sensitive to mental state representation was taken from earlier work (Tamir et al., 2016). Results were combined across participants via t-tests on r-to-z transformed correlations. Error bars indicate 95% confidence intervals from percentile bootstraps of the mean (untransformed) correlation values. Dashed vertical lines separate models derived from different

Figure 2.3 (Continued). procedures: the five models on the left were based on self-report ratings; holistic similarity measures were calculated based on reaction time and choices in a triplet similarity judgment task; and the confidence model was derived from in-scanner ratings following each trial.

Pattern similarity on a trial-wise basis was analyzed within the voxels selected by the independently-defined mask. This multiple regression RSA mirrored the primary searchlight analysis described above, although over a larger spatial scale. Within these areas – previously identified as sensitive to mental state representation (Tamir et al., 2016) – we found that pattern similarity between experimental trials reflected the influence of both target person identity (d = .89, p = .0003) and context identity (d = .78, p = .001). Although not a hypothesis of substantive interest, we also note that the standardized effect of run on pattern similarity was very large (d = 2.84). The substantive effects were very similar in magnitude when heavily smoothed patterns were instead input into the RSA: person identity d = .89, p = .0003; and context identity d = .68, p = .004. These results indicate that the information contained in fine-grained patterns in the searchlight analysis is recapitulated at the coarse spatial scale of inter-regional differentials in the social brain network.

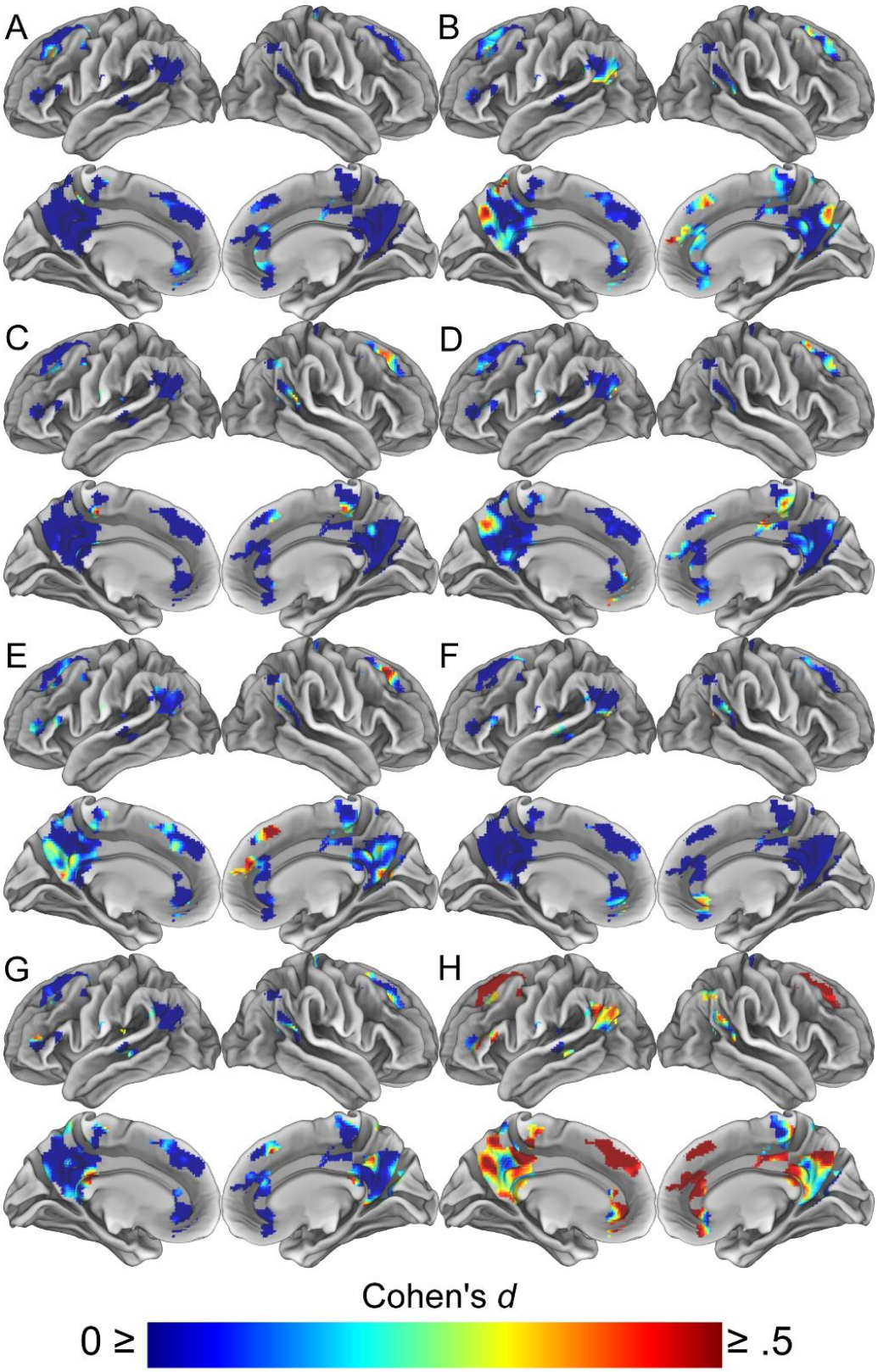
To further probe this result, we attempted to create a descriptive mapping of which voxels that contributed most to the representation of person identity in the smoothed activity patterns. To this end, we repeatedly divided the independently feature-selected regions into 252 random parcels of approximately 100 voxels each. We repeated the person- and context-identity regression RSA using pattern similarity calculated separately across with each parcel of voxels. The regression coefficients could be tabulated on a voxelwise basis across iterations to determine

the typical contribution of each voxel person-encoding coarse-grained patterns. Within each participant, the contribution of individual voxels could be calculated to arbitrary reliability across repeated simulation. The average split-half reliability with respect to voxels across 1000 parcellations was .83. However, the location of these voxels was not at all consistent across participants: the split-half reliability with respect to voxels across 23 participants was -.06. This result suggests that the coarse-grained patterns encoding person identity may rely on specific voxels within each brain, but if so, the location of these voxels is idiosyncratic. In other words, the spatial distribution of person-specific pattern codes appears to be uniformly distributed across the social brain network, at least at the level of the population.

We also computed a voxelwise descriptive mapping of standardized effect sizes for each of the substantive psychological theories tested in the feature-selected RSA. This analysis was conducted by repeating analysis testing these theories exactly as described above, but within the context of a searchlight analysis. The resulting correlation maps reflected the performance of each model in explaining pattern similarity within searchlights centered at each voxel in the brain. The maps were smoothed with a 4 mm FWHM kernel and then combined across participants using the formula for Cohen's *d*. The resulting group maps were masked by the significant results in the primary person-representation searchlight analysis (Figure 1), on the principle that meaningful theory fits could only occur in regions where person-specific patterns actually existed. The results (Figure 2.4) provide an indication of which regions may have particularly contributed to the performance of each theory in the feature-selected RSA.

Figure 2.4. Descriptive mapping of theory effect sizes. Each map reflects the across-participant Cohen's *d* of one of the theories of person perception tested in the feature-selected RSA: (A) social groups, (B) egocentrism, (C) Big 5 personality traits, (D) relational models theory, (E) stereotype content model, (F) holistic implicit similarity, (G) holistic explicit similarity, and (H) confidence. Results reflect a wholebrain searchlight, masked by the voxels found to contain person-specific information in the primary searchlight analysis (i.e., the same voxels used in the reliability-selected RSA). The mapping is descriptive rather than inferential and thus is not corrected for multiple comparisons across voxels.

Figure 2.4 (Continued)



Behavioral results. During the imaging task, participants responded to the vividness probe on 92% of trials on average (*SD* = 8.1%) and responded to the accuracy probe on 95% of trials on average (*SD* = 6.0%). On average, 89% of trials had responses on both items and 98% of trials had responses on at least one item. The high response rates indicate that participants were consistently engaged with the task, especially considering the brief response windows. The mean vividness rating was 3.41 and the average of the standard deviations within each participant was .94. The mean accuracy rating was 3.30 and the average within-participant standard deviation was 1.07. The average correlation between vividness and accuracy was .67 (*SD* = .16), suggesting that these measures tapped related phenomenon, but were not completely redundant. Participants' ratings of targets yielded correlated results across several different conceptions of person perception (Figure 2.5), indicating that these social theories made comparable predictions.

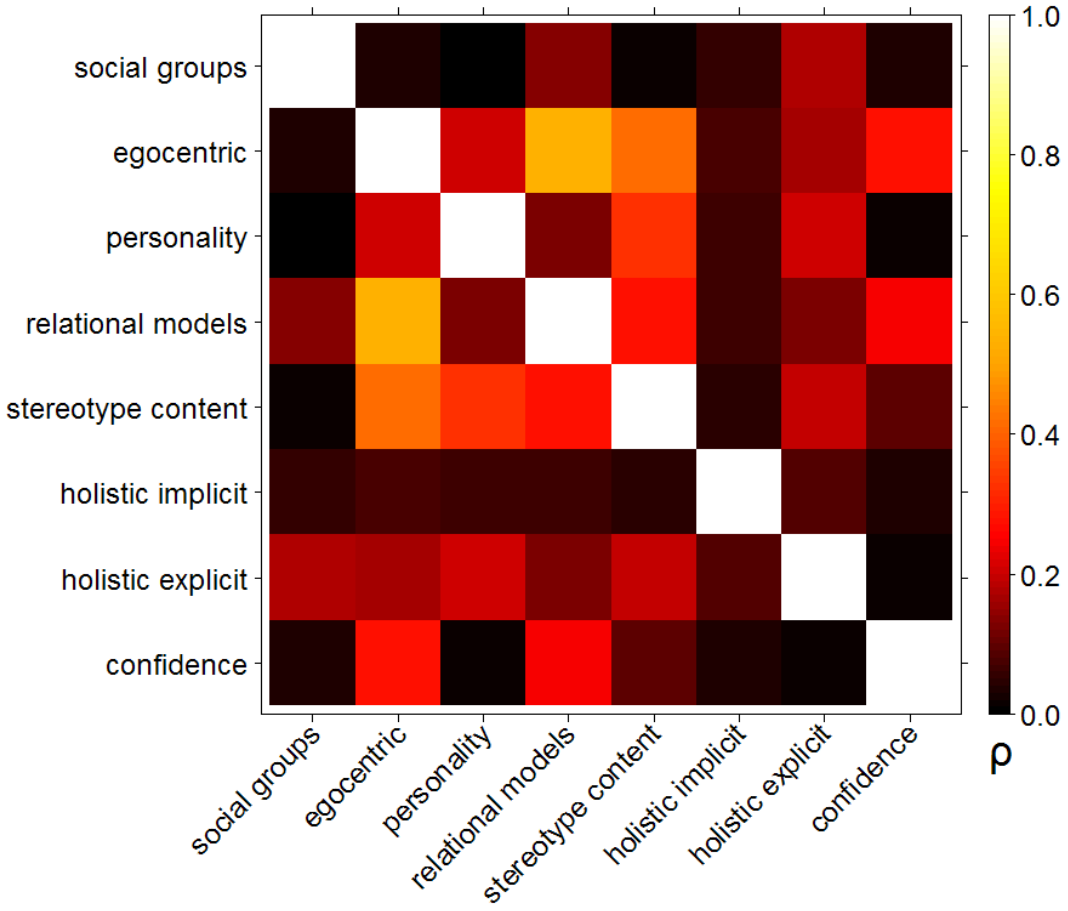


Figure 2.5 *Correlations between potential predictors of neural similarity.* Behavioral and self-report measures of inter-target similarity used in representational similarity analysis were Spearman correlated with each other within each participant. Mean correlations across participants are shown in the heatmap, with larger values in warmer colors.

Discussion

The present study examined the neural representation of personally familiar people. Results of searchlight MVPA indicated that, during mental simulation, target person-specific patterns of neural activity were widely distributed across the social brain network. Person-specific activity patterns overlapped with context-specific patterns in only one area of the brain – medial parietal

cortex – suggesting that this region may play a key role in combining person-specific knowledge with information about context in mental simulations. The central role of medial parietal cortex was further underscored by the fact that portions of this region showed a positive relation between pattern typicality and meta-cognitive perceptions of confidence in simulations of others. This relation suggests that the degree to which person knowledge is integrated into simulations by medial parietal cortex serves as a cue to how confident one should feel about those simulations – that is, how likely they are to predict the course of real world events. Finally, we observed that several measures of interpersonal similarity weakly predicted neural pattern similarity across brain regions that manifest person-specific patterns. The most consistent of these predictors were confidence in simulations (vividness and accuracy) and explicit (self-reported) perceptions of holistic similarity between target people.

The question of whether neural processes and representations are confined to discrete modular regions or are distributed across larger cortical networks has long been of interest to cognitive neuroscientists (de Beeck et al., 2008). The present research considered a particular social cognitive version of this issue: that is, whether person-specific information is widely distributed or confined to a single repository? Although many previous investigations of familiar people have concluded that such knowledge is broadly distributed (Cloutier et al., 2011; Gobbini et al., 2004; Szpunar et al., 2014), a few have emphasized the role of ventral medial prefrontal cortex in particular (Heleven & Van Overwalle, 2016; Welborn & Lieberman, 2015). Our results fall firmly in line with former set of findings.

The current results indicate the presence of person-specific patterns in the social brain network, as defined by sensitivity to differences in others' mental states by a previous study. However, it is worth noting that not all of the regions implicated by our searchlight analysis are

contained within the social brain as commonly defined. For instance, person-specific patterns were also detected within portions of the midcingulate, precentral gyrus, and insula – none of which are generally taken to be socially-specific. Person-specific patterns in medial prefrontal cortex also extended to quite posterior coordinates, overlapping considerably with the anterior cingulate (ACC). Given the ACC's role in error and conflict monitoring (Kerns et al., 2004), this might suggest a role for the ACC in keeping social simulations 'on track' – that is, maximally plausible for individual target people.

The results of this study emphasize the central role that medial parietal cortex, including the precuneus and posterior cingulate, plays in mental simulations of other people. We have observed that 1) this region supports person-specific representations of familiar others, 2) this region also represents context-specific information, and 3) the presence of typical person-specific patterns in this region predicts vividness and accuracy judgments. The overlap between person and context representation hints that this region may contribute to integrating these forms of information, although the present study does not provide direct evidence of this integration. The presence of context representations in medial parietal cortex – and particularly retrosplenial cortex – is highly consistent with previous research (Bar, 2004). However, medial parietal cortex has been associated with diverse mental functions over the history of cognitive neuroscience (for review, see Cavanna & Trimble, 2006), so it seems unlikely that it is specifically devoted to social simulation. The pattern typicality results, however, do implicate it in supporting metacognitive access to such simulations, when they occur.

Representational similarity analysis of person-specific patterns in the social brain network suggested several factors that may shape our mental simulations of other people. Holistic interpersonal similarity among the targets significantly predicted pattern similarity,

suggesting that participant did have conscious access to some of the features contributing to their mental simulations. Confidence (vividness and accuracy) was the overall best predictor of neural pattern similarity between the target people. However, the explanatory success of confidence may be exaggerated by the fact that these ratings were interspersed in the simulation task itself.

Evidence for more specific influences on pattern similarity was weak and inconsistent. Relational models theory (Haslam, 2004; Haslam & Fiske, 1999), which provides a taxonomy of dyadic relationships based on their resource exchange logic, achieved the best performance descriptively but was still only marginally significant within independently selected voxels, and statistically indistinguishable from the other theories tested. If relational models do shape patterns of activity during mental simulation, this would suggest that other people's functional role with respect to ourselves plays an important role in determining how we imagine them acting. We also observed weak evidence in support of the influence of several other theories including the Five Factor Model of personality (Goldberg, 1990; McCrae & Costa, 1987), egocentrism, and the stereotype content model (Fiske et al., 2002). However, the marginal nature of these findings suggests we should be highly tentative in drawing conclusions about which particular factors shape our mental representations of personally familiar people.

Altogether, the present studied used advanced neuroimaging methods to further our understanding of the representation of familiar others. We found evidence for consistent fine-grained patterns of neural activity represent individuals in large sets of personally familiar targets, suggesting that the brain uses a distributed population-coding approach to support person knowledge. Moreover, we found that these person-specific patterns are widely distributed throughout the social brain network, rather than concentrated in a single region. Notably, we also detected coarse-grained person- and context-specific activity patterns which spanned the social

brain network. This suggests that social information encoded at one spatial scale may be recapitulated at other spatial scales. Finally, we found evidence that medial parietal cortex may contribute to the simulation of others' minds in two unique ways: by potentially integrating person and context knowledge, and by providing metacognitive cues to simulation validity.

Paper 3: Theories of person perception predict patterns of neural activity during mentalizing

Thornton, M. A., & Mitchell, J. P. (under revision). Theories of person perception predict patterns of neural activity during mentalizing.

Humans enjoy social lives of tremendous complexity. To successfully navigate this complexity, we must form perceptions of other people. Accurate impressions of others serve the invaluable purpose of allowing us to predict, at least in limited fashion, people's likely thoughts, feelings, and actions. The need to form and use perceptions of others raises the question of how our minds spontaneously organize our perceptions into useful, coherent models of other people. What, if any, implicit psychological taxonomy do we apply to the occupants of our social lives? A popular meta-theory, which spans many existing theories, posits the existence of a representational 'space' within which we meaningfully array the people we know. Theories of this class propose dimensional structures for describing the domain of people. For instance, the stereotype content model (Fiske et al., 2002) hypothesizes that we organize our perceptions of others by evaluating them on two qualities: their warmth and competence. These qualities form the dimensions of a representational space within which the coordinate positions of an individual summarize the person's key social properties. A person or social group's position in stereotype content space thus influences the types of motivations and emotions that perceivers direct toward that person or group.

Such dimensional theories of person perception – models of how we organize our knowledge of others – have been developed to address a number of specific phenomena within social psychology. For instance, the stereotype content model was originally developed for explaining phenomena in an intergroup context (Cuddy et al., 2008). The goal of the present study is to examine whether, and how well, four prominent dimensional theories from the social

and personality literature generalize from their original domains to describe overall perceptions of others. These theories have proven highly successful in dealing with phenomena in their native context, so each might also provide a valuable characterization of general representations of other people. If so, the broader applicability of these theories would allow us to draw on them to explain a much wider range of thoughts and behaviors. In addition, we also examine a fifth, synthetic theory that combined the dimensions of these theories and other social features. This theory represents our attempt to generate and test the best synthesis of existing conceptions of person perception. By doing so, we hope to estimate an upper bound on the field's current explanatory ability. We can also thereby measure the converse – how far our theories have left to go.

Like the stereotype content model, the three other theories we consider were also developed to address relatively specific phenomena. The five factor model of personality – consisting of the "Big 5" trait dimensions of openness, conscientiousness, extraversion, agreeableness, and neuroticism – is a the leading contemporary model of personality (Goldberg, 1990; McCrae & Costa, 1987). As such, it generally aims to capture the objective reality of personality – that is, latent variables that explain individual differences in people's behavioral tendencies – rather than just how personality is perceived by others. Nonetheless, the empirical basis of the five factor model relies extensively on peer perceptions and lay conceptual judgments of trait terms, making it very much a theory of person perception. The third theory we consider is a two factor model of mind perception, consisting of the dimensions of agency and experience (Gray et al., 2007). This theory is based on people's tendency to attribute a mind to various entities. Two factors account for much of this tendency: an entity's perceived capacity for subjective experience and its agentic capacity to influence the world. The fourth and final

model we consider is a two-factor theory of social face perception (Oosterhof & Todorov, 2008). This theory is based on the observation that most of the trait inferences that people make on the basis of viewing others' faces can be reduced to two dimensions: trustworthiness and social dominance. Certain facial features are associated with each dimension: for instance, wide set eyes and upturned mouths drive trustworthiness judgments.

Despite the success of these theories in describing the phenomena to which they were originally tailored, there are also reasons to believe that these theories might not generalize to predicting person perception outside of their original context. The meta-theory of a representational space for other people seems likely to be true at some level, but it is not clear which specific dimensions embody the successful instantiation of meta-theory. The stereotype content model was conceived as an explanation for intergroup affect – will it retain explanatory power when group membership is not at the fore? Do the Big 5 traits characterize actual behavior better than perceptions of others' behavior? Might the dimensions of mind perception matter less when all of the perceptual targets are clearly endowed with minds? Can the dimensions that describe face-based trait inference still shape social judgments when no faces are visible? The answers to any of these questions might easily be "no," in which case the dimensions of the corresponding theory might not be able to serve as the informational basis of mentalizing. Thus, it is far from a forgone conclusion whether the theories we consider will be successful accounts of the mental representation of other people, and if they are, which theories will prove the more or less successful.

Addressing this issue poses a significant challenge due to the diversity of the theories in question. The phenomena, paradigms, and measures applied to these theories thus far have little in common; a new shared framework is needed to compare them directly. Here we import such a

framework, pioneered in other areas of cognitive neuroscience (Huth et al., 2016; T. M. Mitchell et al., 2008), into the social domain. This framework combines functional magnetic resonance imaging (fMRI) and multivoxel pattern analysis (MVPA). This approach has multiple advantages, several of which are due to the nature of fMRI as a measure. First, fMRI is an implicit measure, and thus should minimize the effects of social desirability on outcomes. Second, fMRI is inherently a rich multidimensional measure, because signal is collected simultaneously from voxels throughout the brain. Finally, fMRI permits the use of a wide variety of tasks that lack informative behavioral responses because behavioral responses need not be the primary source of data in an fMRI paradigm.

The modelling approach we adopt here enhances these advantages. Standard fMRI analyses focus on differences in the absolute level of activity across experimental condition, making such analyses well-suited for identifying regions that are activated by a particular cognitive process. This approach of mapping cognitive functions onto brain regions has yielded a wealth of valuable results, such as the discovery that social processes are mediated by neural substrates distinct from those that subserve comparable cognitive processes (J. P. Mitchell, 2009; J. P. Mitchell et al., 2002; Saxe & Kanwisher, 2003; Saxe & Wexler, 2005). However, because few traditional social psychological theories make explicit predictions about the spatial distribution of neural activity, it has been difficult to test such theories directly using standard univariate analyses. In contrast, MVPA – the general term for the techniques we employ – focuses on spatially extended patterns of brain activity across many voxels or regions (Haxby et al., 2001). By analyzing distributed patterns of activity, MVPA can go beyond the examination of *process* to reveal much about the *content* of cognition (Kriegeskorte & Bandettini, 2007; Mur, Bandettini, & Kriegeskorte, 2009). This approach has already borne fruit in social neuroscience,

yielding insight into the mental and neural organization of mental state representations (Skerry & Saxe, 2015; Tamir et al., 2016) and social categories (Stolier & Freeman, 2016). By considering ensembles of voxels together, MVPA is also frequently more sensitive than analogous univariate approaches, even for the detection of univariate signals (Davis et al., 2014). These features make it a natural choice for understanding how the content of person perception is organized.

In the present study, we employ two types of MVPA. The primary analysis we adopt is a form of encoding model that we will refer to as "feature space" or "voxelwise encoding" modelling. In this approach, one or more theories are compared in terms of their ability to predict patterns of brain activity associated with particular stimuli (T. M. Mitchell et al., 2008). In the present case, for instance, a feature space model based on stereotype content would be trained to 'know' what pattern of brain activity is associated with thinking about a warm person, and what pattern is associated with thinking about a competent person. Then, if a new target person is 'introduced' as having specific amounts of warmth and competence, the feature space model can predict what pattern of brain activity the target should elicit by mixing together its canonical warmth and competence patterns in the appropriate ratio. Such predicted patterns can be compared with the actual patterns elicited by the targets to determine the model's accuracy. This accuracy in turn reflects how well the feature dimensions of each theory serve as a basis for describing spontaneous neural activity during mentalizing.

In addition, we supplement this feature space modelling approach with the use of representational similarity analysis (RSA), which assesses the ability of each individual dimension to explain the (dis)similarity between target-specific patterns of brain activity (Kriegeskorte, Mur, & Bandettini, 2008). We also use this approach to test two accounts of pairwise similarity between target people, based on holistic ratings and biographical text

analysis, respectively. Simultaneously, we take advantage of the relative computational simplicity of representational similarity analysis to probe method-related variance in our findings, assessing their robustness to various analytic choices.

By combining the methodological advantages of neuroimaging with these recent innovations in computational modelling, we can directly test five theories of person perception. If the feature space models accurately predict patterns of neural activity elicited by making diverse social judgments about a large set of people, this would suggest that the extant theories generalize well beyond their original purposes and describe the general framework our minds spontaneously deploy to organize representations of others. In addition to learning which of the theories provide better and worse characterization of how we represent people, the current study will allow us to quantify how well important theories perform relative to hypothetical ideal. This uncommon opportunity will help us understand whether we, as a field, are close to solving the problem of person perception or only just beginning.

More broadly, the success of the feature space encoding models would support the meta-theory that we represent others' within a multidimensional representational space. This finding in itself would not be trivial, as a number of alternatives might obtain: for example, an arbitrary pattern might encode each target person, with no relationship between interpersonal similarity and pattern similarity, a variant of the sparse coding or (pattern-wise) "grandmother cell" hypothesis (Gross, 2002). Even if a low-dimensional representational space does exist, the relationship between the dimensions and patterns of brain activity might be highly complex or nonlinear, or encoded at a spatial scale inaccessible to fMRI. Finally, the dimensions of such a space might accord with inscrutably deep computational variables, rather than familiar social dimension to which we have explicit conscious access. If any of these possibilities obtained, it

would invalidate not just one of the specific theories we are testing, but the general meta-theory they together embody.

Methods

Behavioral data, norming data, and processed neuroimaging data, as well as custom analysis and stimulus presentation code, are freely available on the Open Science Framework (https://osf.io/32wrq/). We report how we determined sample size, all data exclusions, all manipulations, and all measures in the study.

Participants. The desired sample size for the neuroimaging experiment was calculated via a resampling-based power analysis using data from a previous study (Tamir et al., 2016). This study was similar in design to the current study, with the same number of stimulus conditions, similar trial durations and counts, and the same fMRI scanner in common. Moreover, it was a rare example of a condition-rich domain mapping study in the social domain, and thus judged more likely to produce realistic effect size projections than other nonsocial studies might. Participants in that study made judgments about the extent to which various pre-tested scenarios would elicit each of 60 mental states such as "embarrassment" or "planning" while in fMRI scanner. Activity patterns associated with each of the 60 states were calculated using the general linear model, and pattern dissimilarity (correlation distance) between these patterns was calculated within regions showing a univariate effect of mental state identity ($p < .0001$, in an omnibus voxelwise ANOVA). Pattern dissimilarities were then correlated with the absolute differences between the mental states on principal components termed rationality, social impact, human mind, and valence. Valence proved the smallest significant predictor of pattern similarity, with an average $r = .05$, and thus we targeted this effect size in the power analysis to be conservative. Simulated samples ranging in size from 10 to 40 were generated by bootstrapping

participants from this data set. Representational similarity analysis was conducted on each participant in the bootstrapped sample, regressing neural pattern similarity between the 60 mental states onto the behavioral predictions of the three orthogonal dimensions. The resulting coefficients were entered into random-effects t-tests for each dimension, and the resulting p-values aggregated across bootstrap samples to estimate power. This procedure is analogous to the representational similarity analyses we conducted in the current study. A target of 30 participants was estimated to provide power greater than .95.

Imaging participants ($N = 30$) were recruited from the Harvard University Psychology Study Pool. One participant was excluded due to failure to respond on 45% of experimental trials (4.7 SDs below mean response rate). The remaining participants (18 female; age range 18-28, mean age = 22.7) were right-handed, neurologically normal, fluent in English, and had normal or corrected-to-normal vision. Two pilot versions of the imaging task (Ns = 51, 45) were conducted outside of the scanner to ensure the functionality of the task and to assist in item selection. Online participants in the stimulus norming surveys (Ns = 316, 648, and 858) were recruited via Amazon Mechanical Turk. All participants provided informed consent in a manner approved by the Committee on the Use of Human Subjects in Research at Harvard University.

Stimuli and behavioral procedure. We applied two criteria in selecting a set of 60 people to serve as mentalizing targets: familiarity and diversity. Familiarity was necessary to ensure that the task was feasible for participants to complete – that is, that they knew the targets about whom they were asked to make inferences. Diversity, within the constraint of familiarity, was desirable to maximize observable effect sizes and ensure the generalizability of our findings to the broader set of targets about whom people actually think in everyday life. To generate a set of targets with these two properties, we employed a two-pass selection procedure. The first pass

used automated Internet-based methods, while the second pass validated and refined the first through the use of large online norming surveys. We began with webscraping and text analyses because these techniques are fully automated, and require no input from human participants. Given the initially large search space, consistently of potentially hundreds or thousands of potential target people, eliciting human judgments (particularly of pairwise similarity) would have been impractical.

Web scraping and text analysis. To select a set of plausibly famous individuals in a minimally biased way we turned to Wikipedia traffic statistics (maintained at www.stats.grok.se). A list of the 1000 most frequently viewed pages (during March 2014) was surveyed for pages about individual people. Fictional characters were excluded. Additionally, Adolf Hitler and Joseph Stalin were excluded because we anticipated that the extreme nature of these individuals might compress the range of judgments made about the others. This process yielded a set of 245 individuals. To confirm that these people had indeed achieved some degree of lasting fame, we then programmatically retrieved their traffic statistics during a separate time period (June 2014). Anyone who achieved at least 30,000 views during this period (1000 per day on average) was retained in the set, yielding 223 individuals.

The next step in the process was to reduce the famous group to a diverse and minimally redundant set. To this end, the text of the remaining individuals' biographies was retrieved using the Wikipedia API for Python (http://pypi.python.org/pypi/wikipedia/). This text was cleaned by the removal of numbers, punctuation, one letter words, and stopwords (content-free grammatical words such as articles and conjunctions). The frequencies of the remaining words were then tabulated and a master list of words used across the person pages was assembled. To eliminate very low frequency words (many of which were, in fact, non-word artifacts) we required that

words included in the master list appear at least twice each in at least two biographies. Additionally, to remove non-discriminative high frequency words, any word that appeared in more than 90% of pages was removed from the master list. The final list consisted of 8260 unique words. The frequency of each word on the master list was calculated for each biography (normalized by the total number of words on that page, prior to the elimination of very high and low frequency words). The semantic "distance" between any pair of pages could then be calculated as the sum of absolute differences in their respective word frequency vectors. To form an ideal list of 150 people for behavioral testing, the person with the greatest semantic distance to all others was first selected. The person with the greatest semantic distance to the target people already selected was then added to the list iteratively until the desired number had been achieved.

Pre-testing. After the completion of the Internet-based stimulus selection, two online surveys were used to validate these measures and further refine the stimulus set. Participants (N = 316) rated how well they knew each of the 150 target people selected in the earlier selection stages on a continuous line scale from 0 (Not at all) to 100 (Extremely). Ratings of 10 or less on this scale were classified as "unknown." We discarded target people who were unknown to at least 25 percent of raters, leaving 73 target people at this stage. In a second survey, participants (N = 648) rated the pairwise similarity of the remaining targets. Each participant would be assigned one reference target and then judge the similarity of all 72 other targets to that person on a continuous line scale from "Very Different" to "Very Similar." At this point we manually removed three more potentially problematic targets: Jeffrey Dahmer, due to the extreme ratings, Steve McQueen, due to the presence two relatively famous people sharing this name, and Anne Frank, due to being the only non-adult on the list. We then sequentially identified the pairs of

people who were rated most similar to each other and eliminated the less famous person in the pair until our final goal of 60 target people had been achieved.

To locate the 60 final targets on the dimensions of the theories of interest, an additional sample of online participants (N = 869) provided norming ratings (Table 3.1). Each participant was asked to rate all 60 targets with respect to only one dimension. They were given a brief description of the dimension in question to maintain consistent definitions across participants. They were then presented with the targets in a random order and asked to rate them on a continuous line scale with anchors appropriate to the relevant dimension. There were thirteen dimensions in total: warmth, competence, agency, experience, trustworthiness, dominance, openness to experience, conscientiousness, extraversion, agreeableness, neuroticism, attractiveness, and intelligence. Although the last two of these dimensions do not together constitute an extant model, we included them because they are extremely widely discussed features of other people in both the scientific literature and lay parlance. Ratings for each target were averaged across participants to provide single scores on each dimension. To exclude participants who did not comply with the task, only those who provided at least 10 unique rating values were included in the composite.

Table 3.1 *The reliability of norming ratings of targets on psychological dimensions.*

Dimension	Theory	*n*	Mean interrater *r*	Cronbach's α
Agency	Mind perception	60	.13	.90
Agreeableness	Five factor personality	58	.31	.96
Attractiveness	None	71	.29	.97
Competence	Stereotype Content	69	.38	.98
Conscientiousness	Five factor personality	65	.41	.98
Dominance	Social face perception	72	.32	.97
Experience	Five factor personality	52	.35	.97
Extraversion	Five factor personality	72	.40	.98
Intelligence	None	65	.52	.99
Neuroticism	Five factor personality	62	.26	.96
Openness	Five factor personality	66	.16	.93
Trustworthiness	Social face perception	64	.39	.98
Warmth	Stereotype Content	65	.30	.97

Dimensionality reduction. To generate an optimal synthetic theory, principal components analysis (PCA) was applied to the correlation matrix between the thirteen rating dimensions described above. Comparison of orthogonal and oblique solutions suggested better fit for correlated component solutions, and thus a direct oblimin rotation was adopted. An optimal solution, as measured by Velicer's MAP and Very Simple Structure (complexity 2), was obtained with three principal components. These components loaded mostly highly onto dominance, warmth, and extraversion (reflected), respectively (Figure 1B). We named them power, valence, and sociality to distinguish them from the rated variables. All of the original dimensions were reasonably well explained by the factor solution, with a mean communality of .88 and a minimum of .71. Component 1 correlated with component 2 at $r = .26$, and with component 3 at $r = .36$ Component 2 and 3 were correlated at $r = .22$. While nontrivial, these correlations were sufficiently small to minimally complicate interpretation, relative to an orthogonal model. The scores from the three factors were used in place of ratings for the purposes of the constructing encoding models for the resulting three-component synthetic theory. Note that the component scores remain virtually identical (rs > .99) if the same PCA is applied to

theoretical dimensions only, excluding the additional dimensions of intelligence and attractiveness.

Figure 3.1 *Dimensions organizing target people.* (A) Correlations between dissimilarity matrices produced by taking the absolute differences in positions of targets on each of the 13 rated dimensions. Two pair-wise measures of (dis)similarity were also included: textual dissimilarity of Wikipedia biography pages, and holistic pairwise ratings. (B) Loadings of the 13 rated dimensions onto the 3 components of the optimal factor solution. Note that the sign of factor solutions is arbitrary, so for name-consistency the direction of the sociality dimension is sign-reversed in later analyses.

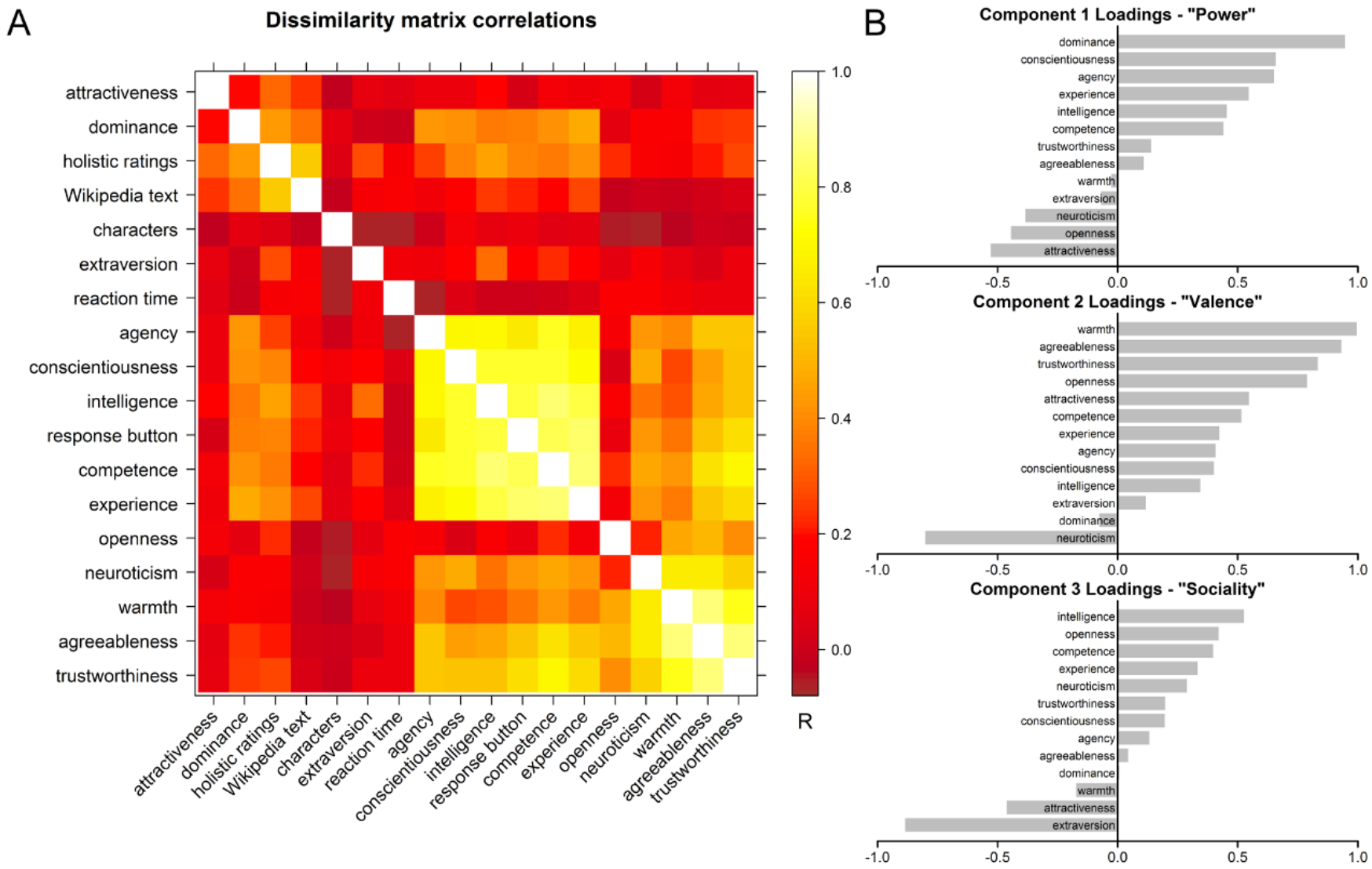


Figure 3.1 (Continued)

Experimental paradigm. The imaging paradigm consisted of a modified version of a mentalizing task used in previous research (J. P. Mitchell, Macrae, et al., 2006; Tamir & Mitchell, 2010, 2013). On each trial, the name of one of the 60 targets would appear in the top center of the screen. After 500 ms, 1 of 12 social judgment items would appear on the screen along with a 5-point Likert type scale. These items consisted of statements such as "likes debating issues with others" and "would grieve the loss of a pet." Participants would use a button box in their left hand to rate how well they believed the statement applied to the target person in question from 1 (not at all) to 5 (very much). We anticipated that participants would generally not know the correct answers to these questions with certainty, but would instead have to make an inference based on their overall knowledge of the target. The item and scale remained on the screen during a 3.25s response window. This period would be followed by a 250 ms minimum fixation period, and a variable jitter fixation period (mean jitter = 1.33s, approximately Poisson distributed in 2s increments). Each run of the experiment consisted of 60 trials: one for each of the target people. The runs were also balanced with respect to the social judgment items, with 5 of each of the 12 appearing in every run. An additional 6s fixation period was allowed at the end of each run to ensure capture of hemodynamic responses from the final trials. Over the course of 12 runs, each target person appeared with each of the social judgment items once, fully crossing these factors.

Prior to entering the scanner, imaging participants also completed two rating tasks. First, they rated themselves on the 12 social judgment items described above. Second, they indicated their liking of, familiarity with, and similarity to each of the 60 target people. These three questions were presented in randomly ordered blocks, with targets randomized within each block. Randomization was conducted independently for each block and for each participant.

Item selection and validation. The 12 items (Table 3.2) used in the imaging paradigm were selected for minimal redundancy from a larger set of 24 items tested in two pilot versions of the behavioral task conducted outside the scanner. The initial set of 24 pilot items was selected manually based on three criteria: 1) applicable to all target people, 2) not widely known (for a fact) for most target people, and 3) variety amongst the items. Using the pilot data, we first calculated the average response for each target person on each of the 24 items. Then we selected items sequentially based on maximal residual variance. The first item chosen was thus the one with the greatest variance across the target people. The second item chosen had the largest residual variance after controlling for the first item. The third item had the largest residual variance after controlling for the first two, and so forth. This approach helped to ensure that each item in the final set of 12 differentiated the target people in a minimally redundant way.

Table 3.2 *Item statistics from imaging task.*

Item text	Mean response	Mean SD of response	Mean reaction time (s)
loves to solve difficult problems	3.24	1.18	1.79
enjoys spending time in nature	3.18	0.97	1.80
enjoys learning for its own sake	3.42	1.10	1.80
likes debating issues with others	3.29	1.09	1.81
thinks that a firm handshake is important	3.41	1.00	1.82
likes to deal with problems alone	3.02	0.97	2.01
prefers to avoid conflict when possible	2.92	1.00	2.02
dislikes travelling by airplane	2.38	0.92	1.86
finds the use of profanity offensive	2.32	1.04	1.95
thinks the wealthy have a duty to the poor	3.17	1.06	1.90
would like to learn karate one day	2.89	0.99	1.83
would grieve the loss of a pet	3.51	0.78	1.75

One potential concern regarding the imaging results might be that the social judgment items chosen for the study could be biased towards eliciting representations more consistent with one theory rather than the others. Although we cannot compare the chosen questions to the entire

domain of hypothetical judgments, we can compare the questions to each of the theories. To this end, we calculated the average response (across participants) to each social judgment item for each target person. We then performed a purely behavioral representational similarity analysis comparing the four extant theories to the social judgment items. For each of the four theories, predictions about similarity between targets were derived by taking the Euclidean distance between the targets on the dimensions of the theory. Similarity estimates were derived from the social judgment items in the same way – that is, by taking the Euclidean distance between targets in the 12-D 'space' defined by the items. The predictions of the theories were correlated with the item response similarity estimates to similar degrees: stereotype content, $r = .69$; social face perception, $r = .65$; five factor personality, $r = .71$; mind perception, $r = .76$. Given the small range of these correlations and the fact that their magnitudes do not covary with model accuracy (below), it seems unlikely that item choice spuriously produced the imaging results. Note that the generally high correlations between the item-space and extant theories are to be expected and desired, as these theories were in general crafted to explain the general principles behind specific interpersonal inferences.

Behavioral data analysis. Responses on the imaging task were analyzed to assess their consistency with previous results in the mentalizing literature. We combined Likert ratings from this task with participants pre-scan ratings of themselves with respect to the same social judgment items to calculate trial-wise self-other discrepancy scores. We analyzed these scores using mixed effects modelling, including fixed effects for similarity, reaction time, and their interaction, random intercepts for participant, social judgment item, and target person, and a random slope for reaction time within participant. Reaction times were mean centered, and trials with reaction times less than 500 ms were excluded from analysis (1.5%). Similarity ratings were

scale-centered. Statistical significance was calculated using the Satterthwaite approximation for degrees of freedom.

Imaging Procedure.

Acquisition and preprocessing. Imaging data were acquired at the Harvard University Center for Brain Science using a 3 Tesla Siemens Tim Trio scanner (Siemens, Erlangen, Germany) with a 32 channel head coil. Functional gradient-echo echo-planar images were obtained from the whole brain using a simultaneous multi-slice imaging procedure (69 interleaved slices of 2mm thickness, TR = 2000ms, TE = 30ms, flip angle = 80°, in-plane resolution = 2.00 x 2.00 mm, matrix size = 108 x 108 voxels, 162 measurements per run). Functional images were preprocessed and analyzed using SPM8 (Wellcome Department of Cognitive Neurology, London, UK) as part of the SPM8w package (https://github.com/ddwagner/SPM8w), and in-house scripts in MATLAB and R.

Data were spatially realigned via rigid body transformation to correct for head motion, unwarped, and then normalized to a standard anatomical space (2 mm isotropic voxels) based on the ICBM 152 brain template (Montreal Neurological Institute). The GLM was used to analyze each participant's data in preparation for MVPA. Each target person was modelled as a condition of interest (60 total) using a boxcar regressor which began on each trial when the name of the target appeared and lasted until the participant responded or the response window ended. The regressors were convolved with a canonical hemodynamic response function and entered into the GLM along with additional nuisance covariates: run means and linear trends, six motion realignment parameters, and outlier time points (defined by the Artifact Detection Toolbox). Regression coefficient maps from this analysis were smoothed with a 6mm FWHM Gaussian kernel to improve inter-participant alignment and increase voxelwise reliability.

To remove the influence of global background patterns from the baseline contrast, each voxel was z-scored across targets prior to MVPA. Note that this z-scoring might introduce a slight dependence between test and training sets in subsequent cross-validation. We could directly measure the bias introduced via this procedure (and any other element of the analysis pipeline) through examination of the permutation testing procedure we used to assess statistical significance. We in fact observed a consistent but minute positive bias, such that the median performance of the permuted encoding models was greater than 0 (Figure 3.3A). However, this baseline shift is of negligible magnitude in comparison with the size of the performance of the actual models, and clearly cannot account for their accuracy. The p-values calculated via permutation testing are inherently adjusted for this bias. Moreover, the representational similarity analyses we conducted are not dependent on dependency between test and training sets (as they do not use cross-validation at all) and these analyses yet still manifest robust effects. It should also be noted that z-scoring could not adversely influence the across-participant or across-data set cross-validation analyses we performed, as the dependency introduced by z-scoring applied only within participant.

Reliability-based voxel selection. Inter-participant reliability in univariate responses to the 60 targets was calculated at each voxel in the brain using the formula for standardized Cronbach's α. This approach is based on earlier stability-based approaches to feature selection(T. M. Mitchell et al., 2008). To select voxels for inclusion in inferential analyses, this measure of voxelwise reliability was combined with pattern similarity reliability. This process proceeded as follows: for each voxelwise threshold between 0 and the maximum voxelwise threshold (in increments of .01) the set of voxels with reliability equal to or greater than the threshold was retained. Patterns of neural activity for each of the 60 targets were extracted from the retained

voxels, and these patterns were correlated with each other to produce neural similarity matrices for each participant. The lower triangular portions of these matrices were then correlated to estimate the reliability of pattern similarity across the halves. After this process had been repeated at each voxelwise threshold, the threshold with the highest pattern-wise reliability was used to select a final set of voxels. This voxel set maximized both the reliability of target-specific patterns across the social brain network and, subject to that restraint, the reliability of person-specific activity within individual voxels.

Note that reliability-based voxel selection is in fact independent of the hypotheses being tested. None of the rating data played a role in this analysis, and so the selection was not artificially biased towards any of the tested theories. Voxels were selected solely on the basis of whether they individually, or as part of a pattern, reliability represented specific target people. In the presence of a real effect, this method of voxel selection will lead to larger effect sizes than using randomly selected voxels due to the reduction of correlation attenuation. However, unlike problematic forms of double-dipping, this technique will not lead to the creation or amplification of spurious relationships. In fact, the increase in effect size in this case only serves to provide a more accurate estimate of the performance which would occur under ideal noise free circumstances. Monte Carlo simulations attesting to these facts are available in the OSF repository. Using the full set of targets and participants for feature selection might introduce a small bias into the cross-validation process described below, but this same bias should affect permuted versions of the models, and thus be easily measurable and discountable for the purposes of significant testing, as described above with reference to z-scoring. Moreover, in the representational similarity analyses described below, we directly compare the reliability-based

feature selection with an independent mask, and find no evidence of bias in the former relative to the latter.

Feature space modelling. Feature space modelling is a form of generative computational encoding modelling of neural activity(T. M. Mitchell et al., 2008), in which feature dimensions are used to predict voxelwise activity and thereby reconstruct distributed patterns across the brain. In the present study, the four theories of person perception discussed above, and the synthetic three component theory generated by PCA, were analyzed and compared through feature space modelling. Three different forms of training and validation were applied to these encoding models to ensure the robust generalizability of our results.

Leave-one-target-out. First, each model was trained separately for each of the 29 participants. Canonical patterns for each dimension were generated by taking the average – weighted by z-scored dimension ratings – of 59 of the 60 targets. These canonical patterns were then multiplied by the dimension ratings of the left-out target person and averaged to predict the actual pattern of neural activity associated with that target. This process was repeated leaving out each target sequentially, with total accuracy for each participant calculated as the average correlation between predicted and actual patterns. In addition to calculating these values within the feature-selected region, we repeated the analysis with 7 brain networks defined by functional connectivity in previous work (Yeo et al., 2011). For each network, we calculated the effect size in terms of Cohen's *d*, allowing for descriptive comparison amongst the networks. The goal of this repetition was to examine whether model fit was indeed better within social brain regions (the default network) than in clearly nonsocial regions (e.g. visual and somatosensory cortices).

For the analyses conducted within the feature-selected regions, confidence intervals around mean performance were obtained via percentile bootstrapping of participants (with

10,000 samples). Percentile bootstrapping was again used to test pair-wise difference in model performance via confidence intervals around mean differences with 100,000 bootstrap samples. The resulting confidence intervals were Bonferroni corrected to account for all pairwise comparisons, and thus statistically significant results pass an uncorrected .005 threshold.

Direct null hypothesis significance testing of model performance (versus chance) was also carried out non-parametrically via permutation-based prevalence testing (Allefeld, Görgen, & Haynes, 2016). For each permutation (1000 total) of a given model, the coordinates of the target people in the dimensional space were randomly shuffled with respect to the corresponding patterns of brain activity prior to starting the cross-validation process. These individual-participant permutations were aggregated into second-level permutations by randomly sampling one of the 1000 permutations for each participant. The minimum values of each of 10^8 such second-level permutations were combined into null distributions, and compared to the actual minimal performance (across participants) from each model. We generated p-values for the global null hypothesis of each model by counting the number of times the permuted statistics were larger than the true minimal statistics (plus one), and then dividing by the number of permutations (plus one). These values allowed us to calculate the lower bound of the 95% confidence interval around the population prevalence of each model's influence. This value can be obtained by subtracting the (N^{th} root of the) p-values described above from the (N^{th} root of the) of the α threshold (i.e., $.05^{(1/29)}$) and dividing the result by one minus the (N^{th} root of the) p-values.

Leave-one-participant-out. The second cross-validation procedure was used to assess generalization across participants rather than across target people. Patterns from 28 of the 29 were averaged to create a single set of 60, and these were then combined via dimension-weighted

average to produce canonical patterns for each of the dimensions of the theories in question. Predicted patterns were then generated for each of the 60 target people based on the canonical patterns and the targets' scores on the dimensions. These predicted patterns were then respectively correlated with the 60 actual patterns from the left-out participant. The pattern reconstruction correlations were averaged to produce a single measure of model performance for that participant, and the process was repeated leaving out each of the remaining participants in turn. The leave-one-participant-out cross-validation scheme also allowed for the estimation of a noise ceiling on performance. To calculate this ceiling, the 60 patterns generated by averaging across 28 participants were directly correlated with the corresponding 60 patterns from the left-out participant without intervention from any dimensional model. The resulting mean pattern correlations indicated the highest meaningful performance a model could achieve, given the inter-participant reliability of the data. Bootstrapping across participants is not appropriate in this case due to the nature of the cross-validation, but permutation testing as described above was again used to perform direct significance testing.

Across data set prediction. The final validation procedure was not cross-validation within the current sample, but rather prediction of a completely independent test dataset. In a previous study of similar design, the authors examined the neural representation of others' mental states (Tamir et al., 2016). Two of the theories tested in the present study – the stereotype content model and the agency and experience model of mind perception – were also examined in that study. Additionally, a synthetic three-factor model emerged from PCA of the ratings of the mental states along 16 theoretical dimensions under consideration. In terms of component loadings, the dimensions of that model – which we named rationality, social impact, and valence – closely resemble the three factors extracted in the present study: power, sociality, and valence

respectively. Using the two extant theories and the 3-component synthetic theory, we tested whether encoding models trained on neural representations of others' traits could reconstruct patterns of activity elicited by thinking about others' mental states.

To achieve this, we extracted patterns for the 60 mental states in the earlier study from within the feature selected regions in the current study. We then averaged patterns across participants – 29 in the current study, and 20 in the previous study – to produce 60 patterns of activity associated with target people and 60 patterns associated with mental states. We generated canonical patterns for the dimensions of the theories in question via weighted-averaging across the target person patterns as described above. We then generated predicted patterns based on recombining the canonical patterns with weights determined by the z-scored dimension ratings of the mental states. Pattern reconstruction accuracy was measured as the mean correlation between predicted and actual patterns. We again calculated noise ceilings for the purpose of comparison, this time based on a leave-one-target-out cross-validation procedure for each model within the mental states data. These ceilings reflect how well each model could predict neural representations of mental states when trained on the same dataset.

Note that this validation approach represents a particularly stringent challenge for the feature encoding models due to the differences between the two datasets. First, completely different sets of participants took part in each study, making inter-individual generalization a necessary condition for successful prediction. Second, numerous imaging parameters – such as spatial and temporal resolution – differed between studies. Third, although both studies involved making inferences about the minds of other people, the mechanics of the tasks were quite different. In the current study, participants made inferences about how well statements applied to well-known targets, whereas in the earlier study participants judged which of two scenarios

would elicit more of a particular mental state in a generic other person. The influence of task differences on activity patterns would likely work against the encoding models, at least by adding noise. Finally, the stimulus-patterns corresponded to people in the present study, but to mental states in the former study. Thus for the encoding models to successfully reconstruct patterns, it would be necessary for a common neural code – correlated with the theoretical dimensions – to represent both people's traits and mental states. The presence of a neural code transcending the trait-state boundary would be a substantial discovery in itself, though in this case it represents but one of many barriers to accurate pattern reconstruction.

Representational similarity analysis. RSA was used to probe the influence of individual dimensions on neural similarity. We employed representational similarity analysis as a complementary approach to the feature encodings in two respects: first, it is much more computationally tractable, allowing for the examination of numerous single dimensions, crossed with several methodological robustness checks, in a reasonable amount of time; second, RSA makes a natural vehicle for the assessment of feature-less models which only describe pairwise differences between stimuli.

The 11 dimensions from the extant theories, as well as the dimensions of intelligence and attractiveness, were tested. In addition to these dimensional accounts, two holistic predictors of pattern similarity were entered into the same analysis. Predictions of dissimilarity between target people were calculated by taking all of the pairwise absolute differences between targets along each dimension in turn. These measures were the average pairwise holistic similarity ratings provided during pre-testing, and the similarity measure computed from the targets' Wikipedia text. These measures were converted to dissimilarities by reflection where necessary. We also included three task features as predictors in the RSA: the average button response and reaction

time to each target, the character lengths of their names. Neural dissimilarity was calculated as the correlation distance between patterns of neural activity for each pair of targets. These neural dissimilarity measures were Pearson correlated with the predictions of dissimilarity described above, considering only the non-redundant lower-triangular elements of each square dissimilarity matrix. Significance testing was conducted directly via permutation testing (N = 1,000) by randomly flipping the signs of the correlation coefficients to generate a null distribution for the average correlation, and indirectly via percentile bootstrapping (N = 10,000) across participants.

In addition to this 'standard' analysis, which mirrored the analytic path of the encoding models as closely as possible, we also examined five variants to assess the methodological robustness of our approach. In the first variant, we replaced the Pearson correlation between predicted and actual dissimilarities with a Spearman rank correlation, which has been recommended for this purpose (Kriegeskorte, Mur, & Bandettini, 2008). Second, we calculated pattern dissimilarity via Euclidean distance rather than correlation distance. Third, we used unsmoothed regression coefficient maps from the GLM, rather than smoothed patterns. Fourth, we used an alternative independent feature selection approach – a mask defined as sensitive to mental state representation in a previous study (Tamir et al., 2016).

Fifth, we controlled for two visual confounds in the imaging task: the length of targets' names, and the feedback provided during the task (i.e., the selected value on the Likert scale lightened slightly from grey to white). We computed the length of targets names directly from character length, and calculated the average slider position for each target within each participant. These values were then converted to dissimilarity values by taking the absolute differences between targets. The resulting RSAs were partial correlations between the theoretical dimensions and neural pattern similarity, controlling for the dissimilarities predicted by the

visual confounds. Note that this procedure is quite conservative, because it assumes that response-related variance in pattern similarity stems from the visual feedback. A plausible alternative might be that responses only correlate with pattern similarity because responses are themselves influenced by social dimensions.

Permutation testing and bootstrapping were repeated for each variant. Additionally, noise ceilings were calculated for each of the method variants, as well as the standard model. This was achieved by averaging the neural dissimilarity matrices of 28 of the 29 participants and correlating the resulting composite with the neural dissimilarity matrix of the left out participant. This process was repeated leaving out each participant in turn, with the final noise ceiling taken to be the average correlation across participants.

RSA was also used to assess the extent of overlap between the four existing theories. In this analysis, the dissimilarity predictions from the dimensions of each theory were fit to the average neural dissimilarity matrix via multiple regression. The fitted values from these regressions were then correlated with each other. This approach effectively uses the neural data to determine the appropriate weighting for combining the dimensions within each theory. The resulting correlations of fitted values reflect the extent to which each of the theories makes the same accurate predictions as the others with respect to neural pattern similarity. The two visual confounds described above were also included to estimate the degree of the confound.

Results

Behavioral results. Participants responded to an average of 96.5% (*SD* = 4%) of trials in the imaging task, indicating high engagement. Behavioral results replicated a number of well-established findings in the literature. Increased similarity between perceiver and target predicted diminished self-other discrepancy (b = -.04, β = -.10, $p < 2 \times 10^{-16}$), as predicted by the

similarity-contingency model (Ames, 2004). Longer reaction time predicted greater self-other discrepancy (b = .12, β = .07, $p < 3 \times 10^{-4}$), replicating findings on egocentric anchoring-and-adjustment in mentalizing (Epley, Keysar, Van Boven, & Gilovich, 2004). Finally, consistent with recent results (Tamir & Mitchell, 2013), similarity moderated the relationship between reaction time and self-other discrepancy, such that anchoring and adjustment was observed more for similar than dissimilar targets (b = .03, β = .04, $p < 2 \times 10^{-5}$). The aforementioned effects remained statistically significant when familiarity and liking ratings were also included in the model. Together, these results provide a clear indication that participants in the imaging experiment were performing mentalizing as previously defined in the literature.

The positions of the target people on the dimensions of four extant theories were established through additional online norming surveys. The idiosyncrasy of such perceptions was reflected in modest average interrater correlations (Table 3.1). However, the composites all ultimately achieved high levels of reliability, with an average Cronbach's α = .96 (SD = .02). We observed large correlations between the dimensions of person perception theories, indicating considerable redundancy between the predictions of the theory dimensions (Figure 3.1A). We conducted principal components analysis (PCA) on the traits to synthesize a single parsimonious theory that encapsulates the unique predictions of all rated dimensions. The three components of the optimal solution loaded most heavily on dominance, warmth, extraversion (Figure 3.1B). To distinguish between manifest and latent variables, we labelled the components power, valence, and sociality, respectively. The components scores were used as the dimensions of a synthetic fifth theory. The synthetic theory thus effectively spanned the representational spaces described by existing conceptions of person perception, allowing us to use it to estimate an upper bound on the explanatory ability of existing research.

Feature selection. The reliability-based feature selection method yielded a set of 10216 voxels for further analysis. The regions selected overlapped substantially with regions previously implicated in social cognition (J. P. Mitchell, 2009; J. P. Mitchell et al., 2002; Saxe & Kanwisher, 2003; Saxe & Wexler, 2005), including medial prefrontal cortex, posterior parietal cortex, the temporoparietal junction, and portions of the lateral temporal lobe (Figure 3.2). This confirms that thinking about different well-known target people elicits reliability differentiable neural activity within the social brain network.

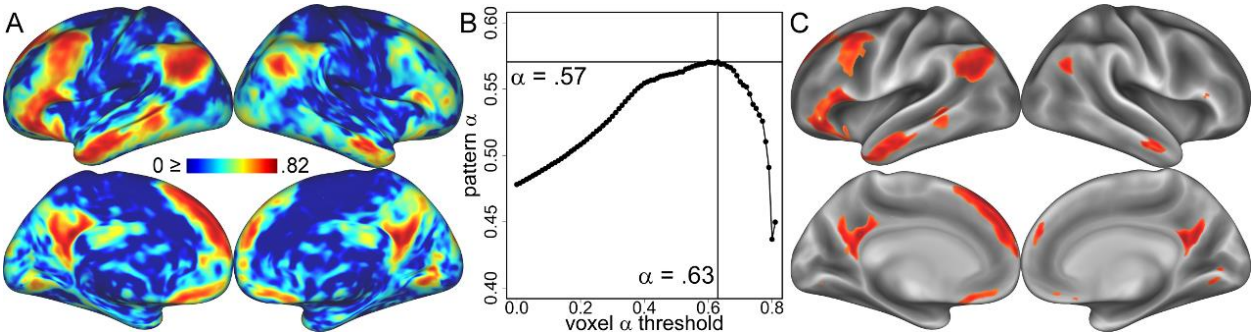


Figure 3.2 *Reliability-based feature selection.* Univariate reliability across targets (A) was calculated for each voxel via the formula for standardized Cronbach's α. The reliability of the correlation matrix between patterns was assessed at each voxelwise reliability threshold between 0 and the observed maximum (B). The voxelwise reliability threshold that maximized pattern correlation reliability was used to select voxels for further analysis (C).

We also examined the reliability of target pattern similarity in a number of a priori ROIs defined from functional connectivity networks in previous research (Yeo et al., 2011). We found that this reliability was highest in the default mode network (α = .54). The lowest pattern reliability was observed in somatosensory (α = .13) and visual cortices (α = .20). Intermediate values were observed for other association cortex networks: dorsal attention – .29, ventral attention – .21, limbic – .24, and frontoparietal – .47. The residual pattern similarity reliability in

the default network – controlling for pattern similarity in the visual and somatosensory cortices – was .49. Together these results support the conclusion that target-specific activity patterns are most pronounced in the default/social brain network, and help to rule out the possibility that visual or motor confounds might account for observed effects.

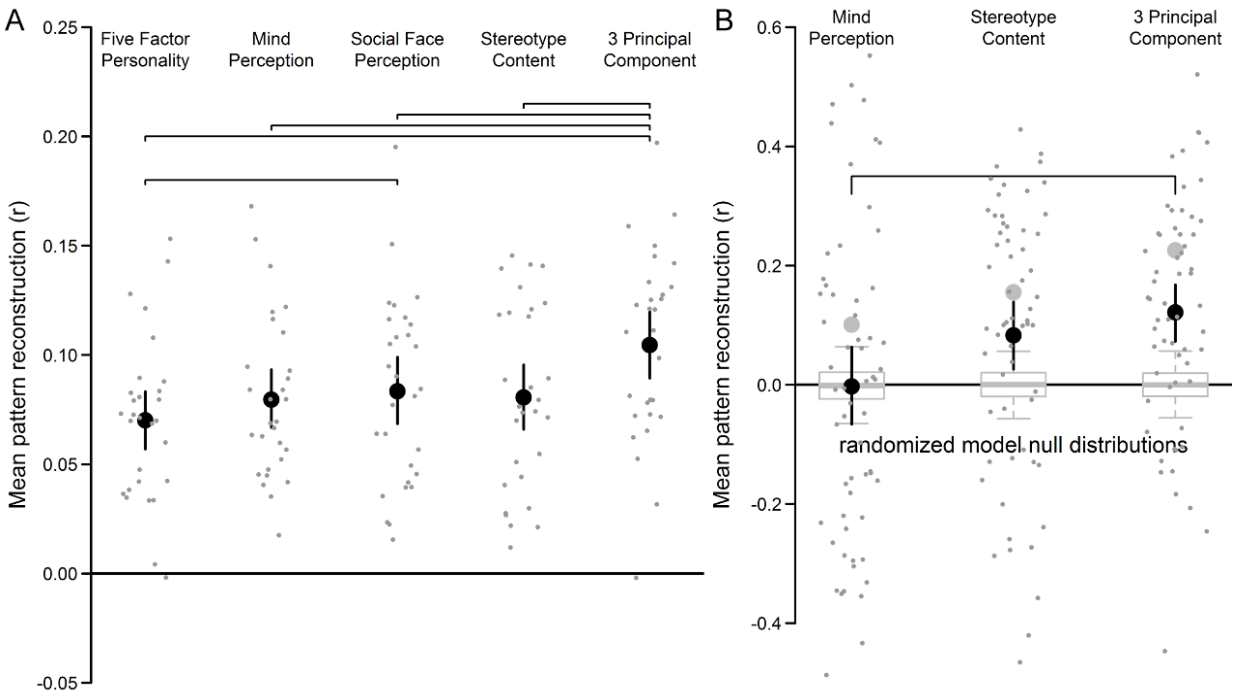


Figure 3.3 *Feature space modelling performance.* Large black circles indicate mean pattern reconstruction, and the error bars around them indicate 95% CIs. Brackets indicate significant pairwise differences. Model performance in leave-one-target-out cross-validation analysis (A) was above chance for all five models. Small grey points indicate the performance of individual participants. Across-dataset pattern reconstruction accuracy (B) was above chance for two of three theories – stereotype content and the synthetic 3 component model. Small grey circles indicate pattern reconstruction accuracy for each of 60 states.

Feature space modelling. In the first cross-validation approach, encoding models were fit separately to each participant's data and tested using a leave-one-target-out procedure, in which performance was measured in terms of the reconstruction accuracy of the model (i.e., the correlation between predicted and actual patterns for the left-out target person). Results reflected statistically significant (ps < .001) performance for all five theories (Figure 3.3A) as assessed by bootstrapping. The lower bound of the 95% confidence interval on population prevalence of each model was 77%, though this figure is likely conservative, as it is determined exclusively by the sample size and number of permutations we could compute in a reasonable time (Allefeld et al., 2016). Average pattern reconstruction accuracies were .070 for the five factor model (Cohen's d = 1.89), .080 for the mind perception model (d = 2.16), .083 for the social face perception model (d = 1.94), .081 for the stereotype content model (d = 1.90), and .104 for synthetic the three principal component model (d = 2.41). The pattern reconstruction values reflect the average correlation between predicted and actual patterns of neural activity for each target person. Model performance (as measured by Cohen's ds) in analogous analyses conducted in brain networks derived from functional connectivity (Yeo et al., 2011), yielded consistent results (Table 3.3). The three component synthetic model achieved the highest performance in all brain networks except the limbic system, where it was slightly outperformed by the stereotype content model. Similarly, model performance was highest in the default network for every model except the stereotype content model, which was better fitted to the limbic system.

Table 3.3 *Leave-one-target-out cross-validation performance (Cohen's d) by brain network.*

Brain network	Five Factor Personality	Mind Perception	Social Face Perception	Stereotype Content	3 Component
Visual	0.94	1.28	1.10	1.30	1.33
Somatosensory	0.71	0.77	1.06	0.95	1.17
Dorsal Attention	1.27	1.52	1.44	1.51	2.00
Ventral Attention	0.74	1.04	0.99	0.86	1.24
Limbic	1.34	1.13	1.06	1.74	1.61
Frontoparietal	1.19	1.32	1.39	1.39	1.68
Default	1.45	2.16	1.73	1.59	2.03

Pairwise indirect difference tests (via bootstrapping) in the reliability selected regions indicated that the three component model significantly outperformed the four extant models, and the social face perception model significantly outperformed the five factor personality model (Table 3.4). Note that the latter outcome demonstrates that more parsimonious theories (the 2-D face perception model) can indeed outperform more complex theories (the five factor model) in the present framework, despite the absence of discounting for higher dimensional theories.

Table 3.4 *Pairwise model comparisons in leave-one-target-out cross-validation.*

Pair	Model 1	Model 2	Bootstrap medians	95% CI Lower Bound	95% CI Upper Bound
1	Five Factor Personality	Mind Perception	-.010	-.027	.009
2	Five Factor Personality	Social Face Perception	-.013	-.027	.000
3	Five Factor Personality	Stereotype Content	-.010	-.022	.001
4	Five Factor Personality	3 Component	-.034	-.044	-.024
5	Mind Perception	Social Face Perception	-.004	-.018	.012
6	Mind Perception	Stereotype Content	-.001	-.018	.014
7	Mind Perception	3 Component	-.025	-.041	-.010
8	Social Face Perception	Stereotype Content	.003	-.011	.018
9	Social Face Perception	3 Component	-.021	-.033	-.009
10	Stereotype Content	3 Component	-.024	-.035	-.012

Values reflect differences in mean pattern reconstruction (*r*) between theoretical models (Model 1 – Model 2). Confidence intervals (CIs) are Bonferroni-corrected for multiple comparisons.

The second validation procedure was leave-one-participant-out cross-validation: the feature space encoding models were trained on patterns averaged across all but one participant and then tested on the excluded participant, iteratively. Again, all five models performed significantly above chance (*p*s < .001). The fact that the encoding models generalized across participants indicates the existence of a common representational topology across brains – that is, the same voxels appear to encode the same dimensions in the brains of different participants. The absolute levels of pattern reconstruction accuracies were somewhat lower in this case then when trained and tested within participant: .063 for the five factor model, .053 for the mind perception model, .056 for the social face perception model, .061 for the stereotype content model, and .072

for the three principal component model. This drop in performance indicates that not all of the reliable variance in neural activity patterns is shared across participants, though this may be merely due to imperfect alignment rather than substantively idiosyncratic coding schemes. Note that we cannot report valid Cohen's *d*s for this analysis because the outcomes of individual participants are no longer independent due to the nature of the cross-validation procedure. It is worth observing that the five factor model – which had the worst performance in the leave-one-target-out case – here outperformed the other three extant models. This hints that the neural topography of the encoding of the Big 5 traits may be more universal than that of, for instance, the social face perception dimensions.

The leave-one-participant-out cross-validation approach allowed for an alternative way of assessing performance: relative to the possible performance attainable given the shared variance across participants. To implement this approach, we divided the raw performance values above by the data's noise ceiling. The results indicated that the five models achieved approximately half to two-thirds of the maximum performance possible given the variance shared across participants: .59 for the five factor model, .49 for the mind perception model, .52 for the social face perception model, .56 for the stereotype content model, and .67 for the three principal component model. We applied the same approach to voxelwise model performance: that is, we calculated correlations between predicted and actual activity across the 60 targets (Figure 3.4) within each voxel of the feature selected region.

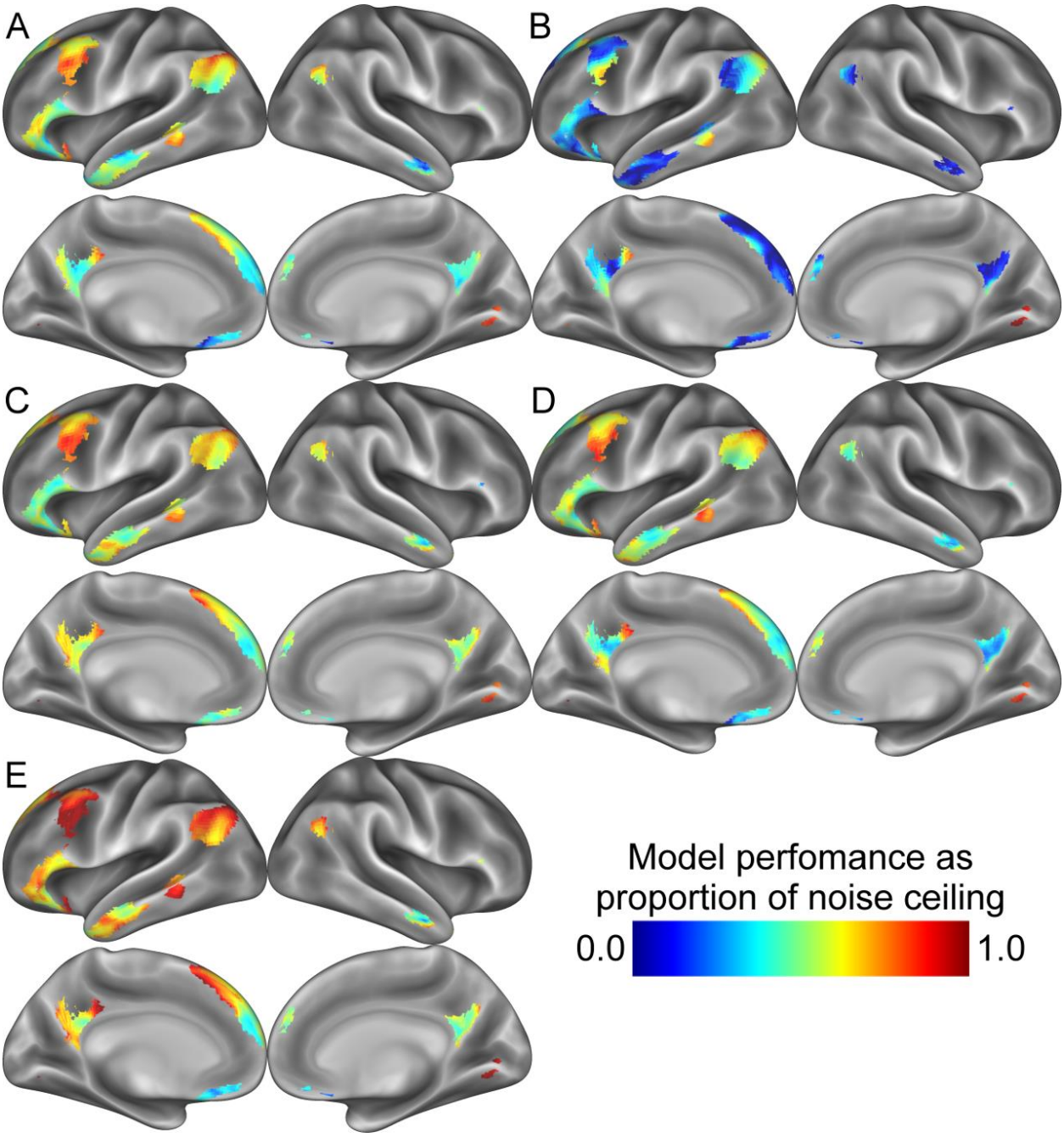


Figure 3.4 *Voxelwise model performance in leave-one-participant-out cross-validation.* Voxelwise correlations between model predictions and actual activity in a leave-one-participant-out cross-validation analysis are shown for each of the five theories under consideration: the five factor model (A), the mind perception model (B), the social face perception model (C), the stereotype content model (D), and the synthetic three-component model (E).

In a third validation method we used a completely independent dataset in a study which probed the neural representation of mental state representation (Tamir et al., 2016). Despite the particularly high barriers to accuracy in this case, two of the three theories successfully reconstructed patterns in the other data set (Figure 3.3B) – the stereotype content model, with a mean pattern reconstruction of .083 ($p < .005$), and the three component model, with a mean pattern reconstruction of .12 ($p < .001$). The mind perception model failed to significantly predict, with a mean pattern accuracy of -.0003 ($p > .1$). Both successful trait-trained theories achieved .54 of their models possible performance by this metric. Pairwise difference testing indicated that the synthetic three component model significantly outperformed the mind perception model. There was considerable variance in the accuracy of pattern reconstruction between different states, but we observed negligible correlations (rs < .1) between accuracy and the dimension of the theories and no obvious pattern in which patterns were most (in)accurately constructed.

The functional topography of the three component model (Figure 3.5) trained on the across-participant averaged data suggests general activation in response to higher power targets, general deactivation in response to higher valence (more positive) targets, and a mixture of activation and deactivation for high sociality targets. However, allowing for these broad directional differences, there was considerable heterogeneity in terms of which regions were more or less associated with each component dimension. Correlating these maps voxelwise with the analogous correlation maps produced from the mental states data set (after Fisher's transforming both), we observed values of r = .43 for power/rationality, r = .28 for valence, and r = .29 for sociality/social impact. This suggests that the dimension of power/rationality is the most conserved across the trait-state divide.

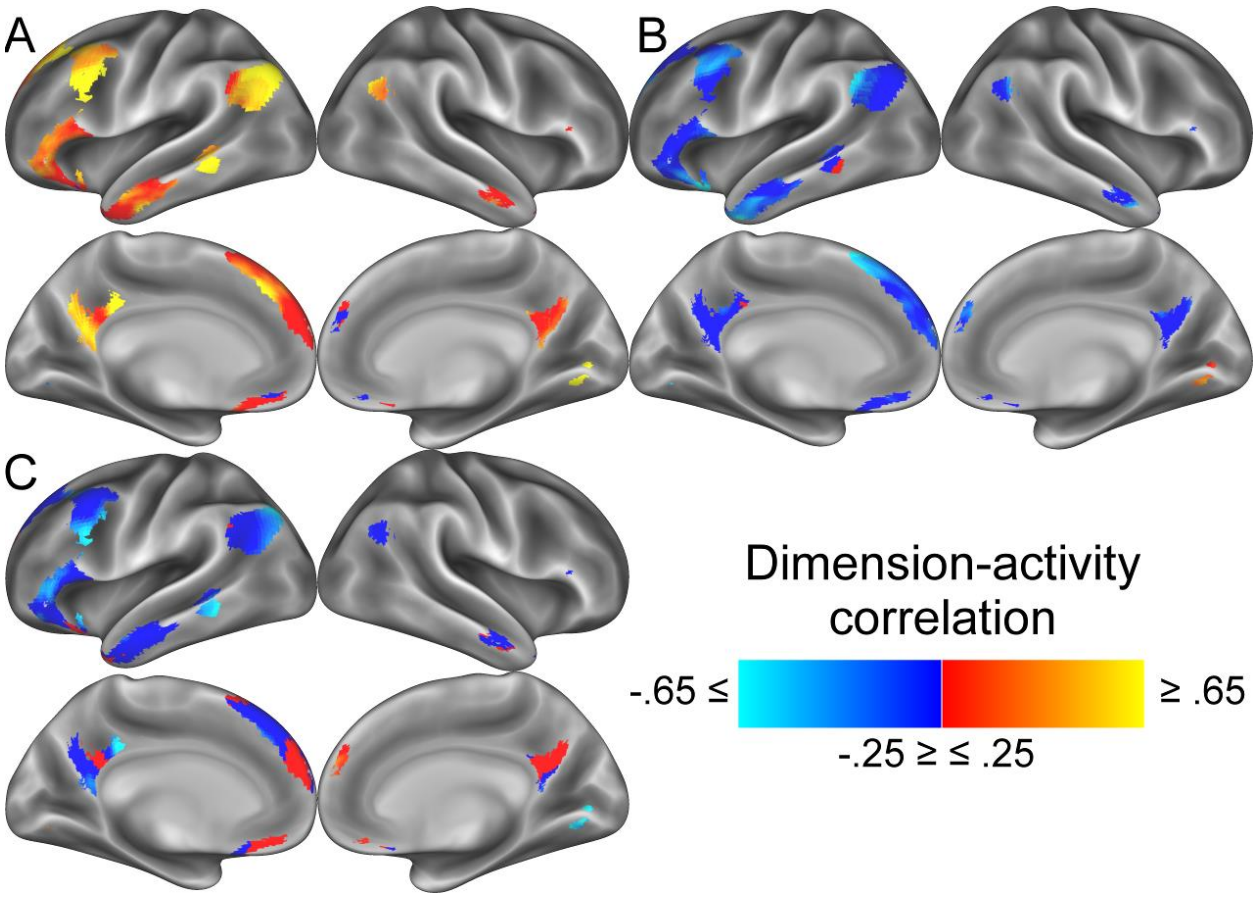


Figure 3.5 *The functional topology of the three-component encoding model.* Colors reflect a voxelwise mapping of correlations between average univariate activity and the three principal components of the synthetic theory: power (A), valence (B), and sociality (C). Activity in orange areas is thus positively associated with thinking about dominant, warm, and extraverted public figures respectively.

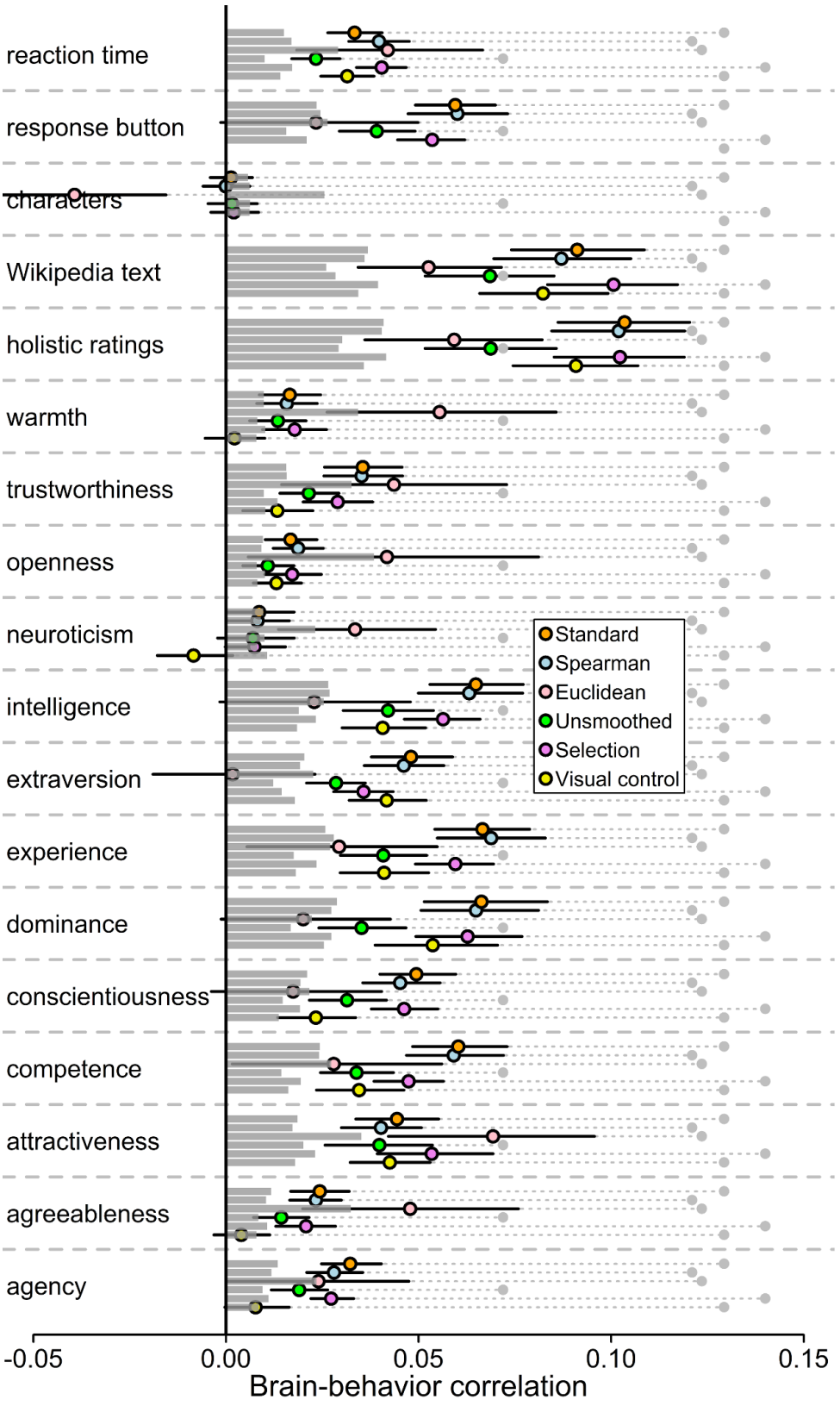
Representational similarity analysis. Representational similarity analysis was conducted to test the ability of individual dimensions to explain the (dis)similarity of patterns of neural activity associated with different target people (Figure 3.6). The two pairwise predictors – holistic ratings and Wikipedia text – clearly achieved the best performance. However, only one psychological dimension – neuroticism – consistently failed to significantly predict pattern

similarity. The task features of average response button (1-5) and reaction time for each target both predicted pattern similarity, but the character length of the target names did not.

Examining the influence of methods also yielded clear results: results for each dimension were highly consistent across variant methodologies. The standard and Spearman analyses performed almost identically in every case. The alternative feature selection also produced results consistent with the standard approach, though sometimes slightly better or worse. The consistency between these features selection methods confirms the unbiased nature of the reliability-based feature selection. The analysis of the unsmoothed data was highly associated with the other methods across dimensions, but was consistently lower in absolute value (though frequently greater as a proportion of noise ceiling). The Euclidean distance metric was the least consistent methodological variant – it yielded highly variable estimates both within dimensions and across models. Controlling for potential visual confounds (response feedback and target name character lengths) resulted in reductions of the brain-behavior relationship across dimensions. Three previously significant dimensions were reduced to non-significance by controlling for these features: agency, agreeableness, and warmth. However, again, most predictors which were significantly related to pattern similarity in the standard analysis remained so, and the overall pattern of performance across models remained highly consistent with the other method variants. This suggests that the performance of the encoding models may be somewhat inflated, but cannot be entirely attributed to these confounding factors.

Figure 3.6 *Representational similarity analysis.* The correlation between neural pattern dissimilarity and fifteen different predictors was assessed under a standard analysis reflecting the feature space encoding models and five reasonable methodological variants. Colored circles indicate the magnitude of the correlations for each predictor, with 95% CIs around them. Grey circles indicate noise ceilings, and grey bars indicate the .975 quantile of permutation distributions.

Figure 3.6 (Continued)



Multiple regression RSA indicated considerable redundancies between the four extant theories, in terms of the (accurate) predictions they made about neural pattern similarity. We observed correlations between fitted values from the substantive models and the fitted values of the putative visual confounds of .61 for the five factor model, .40 for the mind perception model, .83 for the social face perception model, and .45 for the stereotype content model. Combined with the other RSA results reported above, this suggests that the response button makes similar prediction to the extant models. Correlations between fitted values for the extant models ranged from .36 to .69. The predictions of the mind perception model were the most unique, with correlations of only r = .35 with the five factor model, and rs = .36 for both the social face perception and stereotype content model. The five factor model was involved in the two highest correlations between model fitted values: r = .69 with the social face perception model, and r = .63 with the stereotype content model. The social face perception and stereotype content models made moderately similar accurate predictions about neural pattern similarity: r = .47. Together these results suggest that the theories in question attained their similar accuracy for fairly similar reasons.

Discussion

In the present study, we tested how well four prominent theories of person perception (Cuddy et al., 2008; Fiske et al., 2002; Goldberg, 1990; Gray et al., 2007; McCrae & Costa, 1987; Oosterhof & Todorov, 2008)—and one synthetic theory combining their dimensions—predict patterns of neural activity elicited by mentalizing. The primary result was clear: encoding models based on all five of theories performed substantially above chance. Indeed, each encoding model positively predicted the brain activity of nearly every participant. This outcome strongly supports the idea – shared across the tested theories – that people represent others within a

multidimensional social representational space. Moreover, the fact that the theory-predicted neural activity patterns were elicited by making a variety of realistic social inferences suggests that the brain may draw upon a target person's coordinates within the representational space to inform judgments about that target. In other words, the dimensional theories tested here may serve as a substantial part of the informational basis of mentalizing.

The results of three model validation approaches suggest that the theories in question are highly generalizable. Cross-validation indicated that the models could generalize to patterns of activity associated with "unseen" targets. In other words, encoding models trained on the current data set should be able to predict patterns of activity associated with thinking about any other person, assuming that person's coordinates on the relevant dimensions are known. The encoding models also effectively generalized across participants, indicating that the dimensions of the five theories in question are encoded with a common functional topography shared across brains. This finding suggests that the content-based organization of the portions of neocortex involved in social cognition may ultimately resemble that of better-understood regions, such as sensory cortices, which possess highly consistent topographic maps (Kaas, 1997).

The ability of two theories of person perception – the stereotype content model and our synthetic three-component model – to reconstruct activity patterns across datasets also carries significantly implications. This generalization indicates that the neural patterns that encode theoretical dimensions are robust to many low-level differences between datasets, such as participant samples, imaging parameters, and task demands. Overcoming these barriers makes the encoding models far more useful in practice, as they can potentially be applied across many datasets without the need to train the models separately for each sample or participant. Moreover, the encoding models succeeded despite the fact that they were trained on data in which patterns

represented individual people, and then tested on data in which patterns represented mental states. This finding suggests that a common neural code represents both the temporary and enduring features of other people. Thus, the way in which we think about a person who is momentarily feeling a positive state, such as "happiness," may be fundamentally similar to the way we think about a person who manifests a lasting positive trait, such as "trustworthiness." Although the biological reality of personality traits and mental states might differ, the brain appears to represent our lay conceptions of them using at least some of the same dimensions, and encodes those dimensions with similar patterns of brain activity. This finding helps to clarify the degree to which the brain recognizes the trait-state divide (at least at a conceptual level), a topic of longstanding interest to personality psychologists (Mischel, 1968).

The present results also suggest considerable overlap between the extant theories examined. Explicit ratings revealed that the dimensions of all four theories tend to make similar predictions about the (dis)similarity of target people. The respective encoding models produced surprisingly similar levels of overall accuracy, with few and inconsistent pairwise differences in model performance. Moreover, the theories appear to explain activity similiarly in a given brain region. Furthermore, RSA revealed considerable overlap in extent to which the four theories make similar *correct* predictions about brain activity – suggesting that their similar accuracy is not coincidental but instead results from similar insights shared across theories. These results suggest that the study of several different phenomena may have given rise to generally similar theories of person perception. Across all testing approaches, our three component synthesis of the dimensions of the extant theories outperformed all of the original theories, and as stated above, closely approximates – and cross-decodes – a similar model of mental state representation. On this basis, we suggest that people may consistently consider three

fundamental qualities of other people: the likelihood of another person interacting with them in a significant way (sociality/social impact), the ability of that person to enact their will (competence/power), and the likelihood that they will be good or bad (valence). Determining others' coordinates within such a three-dimensional space would make it possible to engage in a wide range of adaptive behavioral responses.

The performance of the synthetic principal component encoding model could be compared to noise ceilings in two of the three cross-validation procedures we conducted (across participants and datasets). Dividing the average pattern reconstruction of the synthetic theory by these noise ceiling indicates how well the theory does in comparison to a hypothetical ideal theory. Since the synthetic theory combines the components of the four theories we considered, as well as the dimensions of intelligence and attractiveness, its accuracy offers a rough estimate of the overall explanatory power of the person perception literature. The results thus indicate that our modern theories predict the content of person perception about half to two-thirds as well as a hypothetical ideal model. Whether this outcome is encouraging or discouraging depends on one's perspective, but we suggest that it is quite positive, given the nascent state of computational approaches in social psychology. In comparison, considerably more complex computational models have been applied to visual cortex with less success (Khaligh-Razavi & Kriegeskorte, 2014). What an ideal theory of person perception will look like remains unknown – it may simply require the addition of more dimensions, or it may require the relaxation of the assumption that linear dimensions characterize the entire representational space. Application of structure-discovery algorithms (Kemp & Tenenbaum, 2008) to condition-rich datasets may help to address such questions.

Notably, all encoding models in the present study relied on a strong assumption: that psychological dimensions are encoded by the linear activation of a single canonical activity pattern. This is among the simplest imaginable population coding schemes, and yet despite this naïve assumption, we observed remarkably robust performance. It is possible that modelling based on more complex assumptions – for example, bipolar genuinely dimensions, or non-linear mappings between activity patterns and psychological dimensions – might produce even better performance. However, we eschew such elaboration for the moment, both because more complex techniques are commensurately harder to interpret, and because they are prone to overfitting, especially with the limited amount of data available to us at present. The fact the models provide quite accurate despite their simplicity carries implications for our understanding of how the brain encodes social knowledge. Specifically, this result suggests that additive patterns of brain activity, distributed across the social brain network in relatively coarse activity patterns, are linearly related to recognizable dimensions from psychological theories. Arbitrary "pointer" patterns for individual people may still exist in the brain – perhaps as finer spatial scales, such as within the hippocampus – but a sparse coding account is not supported by the present findings.

The results of the present study provide support for the conclusions of a number of previous investigations. A recent fMRI investigation found that the Big 5 dimensions of agreeableness and extraversion could be decoded from portions of the social brain network, findings we replicate using RSA (Hassabis et al., 2014). In their investigation of hippocampal brain activity, Tavares et al. (2015) found evidence for a social representational space based on the dimensions of "power" and "affiliation." Although we do not observe reliable target-related signal in the hippocampus itself, we do observe that the conceptually similar dimensions of

competence and warmth from the stereotype content model quite accurately predict patterns of brain activity in the limbic system more generally (Table 3.3). Purely behavioral results from our dimension pre-testing studies also support recent research that suggests that valence can be separated into social and moral components (Brambilla & Leach, 2014; Goodwin, 2015).

Our examination of method-related variance in the present study suggests that the conclusions we draw are quite robust to reasonable alterations of our analytic approach. However, the relative importance of the various trait dimensions we consider still likely depends in part on the particular stimuli and task used in this study. Making inferences about others' thoughts, opinions, feelings, and preferences is a fairly general social process, but is certainly not representative of all possible contexts in which person perception might occur. Use of more naturalistic stimuli and tasks would serve to expand the conclusions one might justifiably make from this research. Furthermore, even with fairly elaborate stimulus selection techniques, the constraint of using well-known target people places severe strain on the representativeness of the targets. Replicating this work with alternative sets of target people will be crucial in ensuring our results extend to the broader set of potential real-world mentalizing targets.

The visual confound of character name length and response button feedback (the selected response lightened slightly when chosen) place an additional limitation on the interpretation of the present findings. These variables account for the influence of three psychological dimensions on pattern similarity, suggesting that the performance of the encoding models may be somewhat inflated. However, the effect of character length was virtually zero, indicating that only the response button shares variance with both pattern dissimilarity and psychological dimensions. Since the purely visual character length confound had no effect, this hints that the ambiguous response button effect – which could be mediated visually or socially – is more likely driven by

meaningful social considerations than by spurious visual differences. Ultimately more research is needed to decouple the social properties of target people from low-level task features.

Despite these limitations, this study offers useful insight regarding the effectiveness of several theories of person perception and demonstrates a new method for directly testing social cognitive theories using neuroimaging. This approach holds considerable promise as a way to compare ideas from diverse areas of the literature which would otherwise be difficult to integrate. By exploring these domains within a common framework, we may discover a great deal about the regularity (or lack thereof) of social representations in the brain. More generally, we believe that use of such approaches will help move the field in a more cumulative, theory-driven direction and enable more fruitful interaction between psychology and neuroscience. Despite assertions (Newell, 1973) that one cannot "play 20 questions with nature and win," social and personality psychologists appear to have assembled a set of remarkably robust and generalizable theories regarding the content of person perception. By our current estimates, they are indeed more than half way to "winning" this particular game.

General Discussion

The papers presented in this dissertation provide a new perspective on one of humanity's oldest questions: how do people understand one another? This new perspective comes from directly interrogating the brain to determine the principles upon which it organizes social knowledge. Across studies, this approach consistently reveals that the social brain network reliably encodes rich information about the mental states and traits of other people. The brain appears to represent this information at a number of spatial scales, from single voxels millimeters across to patterns that span the full length and breadth of the brain. Critically, the ways in which the brain organizes this social information reflect many influential modern psychological theories of mental state representation and person perception (Cuddy et al., 2008; Goldberg, 1990; Gray et al., 2007; Haslam, 2004, 2006; Oosterhof & Todorov, 2008; Russell, 1980).

Although the psychological dimensions of existing theories do not yet fully explain the representational space of other minds, they are remarkably close in some cases. Three dimensional models synthesized from existing theories proved capable of explaining nearly half to over two-thirds of the variance underlying mental state and famous people representation, respectively. The performance of these low-dimensional models in the social domain stands in contrast to the performance of considerably more complex deep learning algorithms in vision, which frequently fail to achieve equivalent performance (Khaligh-Razavi & Kriegeskorte, 2014). Considering the relatively nascent state of computational approaches to social cognition, the efficacy of such parsimonious theories represents a promising beginning towards a complete taxonomy of state and person representation. For instance, exploratory analyses in Paper 1 suggest that an additional 2-3 dimensions may suffice to generate a near-complete explanation for the neural organization of mental state concepts.

Searching for these missing dimensions should doubtless be a focus of future research. Data from the studies presented here may prove useful in generating hypotheses about what they might be. For example, in Paper 3 I observed highly reliable person-related signal in vMPFC, but none of the existing models could adequately explain this signal. Prior work on vMPFC suggests it plays a key role in social valuation (Hare et al., 2010; Zaki et al., 2014; Zaki, Schirmer, & Mitchell, 2011), suggesting that a missing component of extant person perception theories may be some value component, separate from dimensions like warmth and competence. Unfortunately, reverse-engineering psychological dimensions from brain activity is a far from trivial undertaking. Many algorithms for doing so – such as MDS – yield dimensions oriented at arbitrary angles (e.g., "north" may face "east," and so forth). As the dimensionality of the representational space starts to grow beyond two, this arbitrariness makes it increasingly difficult to identify the canonical orientations of residual dimensions or to give them useful names. There is no guarantee that all dimensions discovered in this way could be assigned a name from an existing list of terms in psychological science – some dimensions may instead represent inscrutably deep computational variables that will only become meaningful as processing theories advance. Nevertheless, putting the brain "in the loop" when it comes to both hypothesis generation and theory testing promises to augment behavioral measures by providing a rich, multidimensional, and largely implicit measure of how the mind organizes social information.

The general approach developed over the course of this dissertation uses functional neuroimaging to directly test psychological theories. Refining this procedure represents an important step towards addressing a common critique of social cognitive neuroscience: that it provides relatively little incremental value over traditional social psychological approaches, relative to the cost of fMRI (Todorov, Harris, & Fiske, 2006). Initial fMRI studies of social

cognition focused heavily on brain mapping – identifying which regions could be associated which functions. This was arguably a necessary step to take before advancing to more refined neuroscience, but it offered limited insight into outstanding issues in social psychology because few existing psychological theories made testable predictions about brain activity. The approach adopted here represents one way to circumvent this problem: rather than testing (potentially nonexistent) predictions about how stimuli should activate particular brain regions, we instead test predictions about the similarity between those stimuli by comparing corresponding patterns of activity. This allow us to test social psychological theories that make no reference to "MPFC" or "TPJ" by instead focusing on whether the dimensions of these theories predict the similarity between, or allow for the reconstruction of, patterns of neural activity. Complementary feature selection tools, such as the reliability-based approach demonstrated in Paper 3, even allow researchers to be quite agnostic (*a priori*) about where in the brain relevant patterns reside. Thus neural mapping of social domains, as demonstrated in the papers above, allows social cognitive neuroscience to contribute directly to the progress of cumulative, theory-driven psychology, of the type now often envisioned (Yarkoni, Poldrack, Van Essen, & Wager, 2010).

The dimensions of the synthetic three dimensional models derived in Papers 1 and 3 bear a considerable surface resemblance to one another: rationality, social impact, and valence in the mental state domain, and power, sociality, and valence in the person domain. As we speculate above, these names may be proxies for variables crucial to evaluating other people: whether a person is likely to affect me (sociality/social impact), whether that effect would be good or bad (valence), and whether that effect would be well or poorly executed (rationality/power). Testing this adaptive account of these dimensions is a high priority for future research.

Combining data from Papers 1 and 3 suggests that the resemblance between the representational spaces of mental states and traits is more than skin-deep: a voxel's sensitivity to *state* rationality is correlated with its sensitivity to *trait* power, and to a lesser degree, similar patterns encode *state* social impact and *trait* sociality, and *state* and *trait* valence. In other words, the pattern of brain activity elicited by thinking about a person currently engaged in a momentarily "rational" state, such as planning, resembles the pattern of brain activity elicited by thinking about person with permanently "powerful" trait, such as competence. Together, the three dimensional theories support an encoding model capable of crossing the trait-state boundary by accurately reconstructing patterns of activity associated with mental states based only on person-specific training data. This result suggest that, at the conceptual level, the brain may only partially recognize distinction between momentary mental states and enduring personality traits, helping to resolve a question of lasting interest in the psychological literature (Mischel, 1968).

The conservation of conceptual dimensions across domains is mirrored by the conservation of neural encoding schemes across brains. In Paper 3, I observe that encoding models of person perception robustly generalize across the brains of different participants. In other words, training a model on one person's brain activity during mentalizing allowed us to reconstruct what another person's brain activity looked like when mentalizing about the same target people. Robust cognitive maps – like retinotopy – have long been observed in better understood brain areas such as early visual cortex (Kaas, 1997). The existence of such maps, and their general layouts, are conserved across virtually all humans. However, until recently it has been unclear whether such spatially organized, developmentally conserved maps exist within areas of association cortex such as the social brain network. Our findings join other studies now

suggesting that such conserved social cognitive maps do indeed exist (Bahnemann, Dziobek, Prehn, Wolf, & Heekeren, 2010; de la Vega, Chang, Banich, Wager, & Yarkoni, 2016).

A number of other neuroanatomical findings generalize across the papers presented here. Results in all three papers manifest a bias towards the left hemisphere. This may result from the consistent use of lexical stimuli across studies, but it is worth examining this question directly. Another consistent result is the engagement of medial parietal cortex, which displays the most reliable signs of social information across all three studies. Again, this may partially result from artifacts, such as the location of this region far from sinuses which contribute to signal distortion. However, the results of Paper 2 suggest two specialized roles for MPC. First, this is the only region in which representation of person-specific and context-specific information overlapped, suggesting that MPFC might play a crucial role in integrating the two. Second, the typicality of person-specific patterns in MPC predicted subsequent judgments of confidence in mental simulations. This hints that the representations in MPC may be used for the metacognitive assessment of mentalizing.

Although the papers in this dissertation have explored three major domains of social knowledge, many more remain. Paper 2 takes the first steps towards investigating how the brain represents context, but a more systematic approach is needed to evaluate the bevy of situational taxonomies recently developed (Rauthmann, 2015). Paper 2 also offers some consideration of dyadic relationships, albeit only in the egocentric case of the participants' relationships with others. A more general understanding of how the brain represents different types of social relationships, from friends and families to rivals and acquaintances, must eventually be sought. Insofar as the ultimate goal of social cognition must be predicting others' behavior, humans must also represent socially relevant actions in a systematic way. Developing and testing an action

taxonomy is thus also a necessary component of any complete theory of social knowledge. In all domains, future research might benefit from taking more naturalistic approaches to sampling from the domain and eliciting brain activity in the scanner. The increasing ubiquity, capability, and affordability of wearable cameras make them ideal tools for achieving ecologically sound domain-sampling. Participants wearing such cameras could passively record hours of their real-world experience in a minimally obtrusive way. Allowing for reactivity due to recording, the people, situations, and actions that populated these recordings would otherwise be highly representative of the participants' natural experiences. This would help to mitigate the challenging problem of bias in the selection of such stimuli. Wearable cameras could also be hybridized with another naturalistic paradigm – viewing movies in the scanner (Hasson & Malach, 2005) – to elicit patterns of brain activity closely approximating the everyday experiences otherwise inaccessible to fMRI. Naturalistic viewing (of movies not personally recorded) has already been combined with fMRI to segment human experience based on brain activity (Baldassano et al., 2016), a clear cousin to understanding the representation of situations. Groups and networks are also key aspects of human experiences, but fortunately some of the first steps towards understanding how the brain organizes these domains have recently been taken by others (Parkinson, Kleinbaum, & Wheatley, 2017; Stolier & Freeman, 2016).

As we begin to understand how the brain makes sense of the social world, it will become increasingly important to consider how the brain uses this information. One influential notion is that the brain in general (Friston, 2010), and the social brain in particular (Kilner, Friston, & Frith, 2007; Koster-Hale & Saxe, 2013), is particularly oriented toward the goal of prediction. This predictive coding hypothesis suggests that, in representing the world, the brain reflexively represents predictions about the future. Making efficient, accurate predictions allows an

organism to anticipate the future and act accordingly, increasing its adaptive fitness. How might the social representational spaces explored here fit into such a framework? In the case of mental states, proximity within such a space might predict the likelihood of transitions between different thoughts and feelings. Thus, knowing that someone is currently feeling angry could allow me to predict that they may be more likely to later feel regretful (perhaps as the result of taking some unfortunate action) because these states are close along at least one dimension of the space (valence). This probabilistic account may be generalized across domains as well: for example, a person with a certain trait, or in a certain state, may be more likely to conduct a certain action in a certain situation. Understanding the particulars of how social knowledge facilitates social prediction will surely be one of the next major challenges in social cognitive neuroscience.

References

Allefeld, C., Görgen, K., & Haynes, J.-D. (2016). Valid population inference for information-based imaging: From the second-level t-test to prevalence inference. *NeuroImage, 141*, 378-392.

Ames, D. R. (2004). Strategies for social inference: A similarity contingency model of projection and stereotyping in attribute prevalence estimates. *Journal of Personality and Social Psychology, 87*(5), 573-585.

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience, 7*(4), 268-277.

Anzellotti, S., Fairhall, S. L., & Caramazza, A. (2013). Decoding representations of face identity that are tolerant to rotation. *Cerebral Cortex, 24*(8), 1988-1995.

Bahnemann, M., Dziobek, I., Prehn, K., Wolf, I., & Heekeren, H. R. (2010). Sociotopy in the temporoparietal cortex: Common versus distinct processes. *Social Cognitive and Affective Neuroscience, 5*(1), 48-58.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2016). Discovering event structure in continuous narrative perception and memory. *bioRxiv*, 081018.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5*(8), 617-629.

Bettencourt, K. C., & Xu, Y. (2013). The role of transverse occipital sulcus in scene perception and its relationship to object individuation in inferior intraparietal sulcus. *Journal of Cognitive Neuroscience, 25*(10), 1711-1722.

Blakemore, S.-J., Winston, J., & Frith, U. (2004). Social cognitive neuroscience: Where are we heading? *Trends in Cognitive Sciences, 8*(5), 216-222.

Brambilla, M., & Leach, C. W. (2014). On the importance of being moral: The distinctive role of morality in social judgment. *Social Cognition, 32*(4), 397-408.

Britton, J. C., Phan, K. L., Taylor, S. F., Welsh, R. C., Berridge, K. C., & Liberzon, I. (2006). Neural correlates of social and nonsocial emotions: An fMRI study. *NeuroImage, 31*(1), 397-409.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904-911.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network. *Annals of the New York Academy of Sciences, 1124*(1), 1-38.

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences, 11*(2), 49-57.

Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: A review of its functional anatomy and behavioural correlates. *Brain, 129*(3), 564-583.

Cloutier, J., Kelley, W. M., & Heatherton, T. F. (2011). The influence of perceptual and knowledge-based familiarity on the neural substrates of face perception. *Social neuroscience, 6*(1), 63-75.

Cohen, M. A., Alvarez, G. A., Nakayama, K., & Konkle, T. (2016). Visual search for object categories is predicted by the representational architecture of high-level visual cortex. *Journal of Neurophysiology*.

Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y.-C., . . . Haxby, J. V. (2012). The representation of biological classes in the human brain. *The Journal of Neuroscience, 32*(8), 2608-2618.

Corradi-Dell'Acqua, C., Hofstetter, C., & Vuilleumier, P. (2014). Cognitive and affective theory of mind share the same local patterns of activity in posterior temporal but not medial prefrontal cortex. *Social Cognitive and Affective Neuroscience, 9*(8), 1175-1184.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, 29*(3), 162-173.

Cuddy, A. J., Fiske, S. T., & 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.

Davis, T., LaRocque, K. F., Mumford, J. A., Norman, K. A., Wagner, A. D., & Poldrack, R. A. (2014). What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. *NeuroImage, 97*, 271-283.

de Beeck, H. P. O., Haushofer, J., & Kanwisher, N. G. (2008). Interpreting fMRI data: Maps, modules and dimensions. *Nature Reviews Neuroscience, 9*(2), 123-135.

de la Vega, A., Chang, L. J., Banich, M. T., Wager, T. D., & Yarkoni, T. (2016). Large-scale meta-analysis of human medial frontal cortex reveals tripartite functional organization. *The Journal of Neuroscience, 36*(24), 6553-6562.

Decety, J., & Grèzes, J. (2006). The power of simulation: Imagining one's own and other's behavior. *Brain Research, 1079*(1), 4-14.

Decety, J., Jackson, P. L., Sommerville, J. A., Chaminade, T., & Meltzoff, A. N. (2004). The neural bases of cooperation and competition: An fMRI investigation. *NeuroImage, 23*(2), 744-751.

Dunbar, R. (1998). The social brain hypothesis. *Brain, 9*(10), 178-190.

Epley, N., Keysar, B., Van Boven, L., & Gilovich, T. (2004). Perspective taking as egocentric anchoring and adjustment. *Journal of Personality and Social Psychology, 87*(3), 327-339.

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.

Fletcher, P., Frith, C., Baker, S., Shallice, T., Frackowiak, R., & Dolan, R. (1995). The mind's eye—precuneus activation in memory-related imagery. *NeuroImage, 2*(3), 195-200.

Forstmann, M., & Burgmer, P. (2015). Adults are intuitive mind-body dualists. *Journal of Experimental Psychology: General, 144*(1), 222-235.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127-138.

Gallagher, H. L., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *NeuroImage, 16*(3), 814-821.

Garavan, H., Pendergrass, J. C., Ross, T. J., Stein, E. A., & Risinger, R. C. (2001). Amygdala response to both positively and negatively valenced stimuli. *Neuroreport, 12*(12), 2779-2783.

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin, 117*(1), 21-38.

Gobbini, I. M., Leibenluft, E., Santiago, N., & Haxby, J. V. (2004). Social and emotional attachment in the neural representation of faces. *NeuroImage, 22*(4), 1628-1635.

Goldberg, L. R. (1990). An alternative "description of personality": The big-five factor structure. *Journal of Personality and Social Psychology, 59*(6), 1216-1229.

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science, 24*(1), 38-44.

Gorno-Tempini, M. L., & Price, C. J. (2001). Identification of famous faces and buildings. *Brain, 124*(10), 2087-2097.

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science, 315*(5812), 619.

Gross, C. G. (2002). Genealogy of the “grandmother cell”. *The Neuroscientist, 8*(5), 512-518.

Hare, T. A., Camerer, C. F., Knoepfle, D. T., O'Doherty, J. P., & Rangel, A. (2010). Value computations in ventral medial prefrontal cortex during charitable decision making incorporate input from regions involved in social cognition. *The Journal of Neuroscience, 30*(2), 583-590.

Haslam, N. (2004). *Relational models theory: A contemporary overview*. London, United Kingdom: Psychology Press.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and social psychology review, 10*(3), 252-264.

Haslam, N., & Fiske, A. P. (1999). Relational models theory: A confirmatory factor analysis. *Personal Relationships, 6*(2), 241-250.

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., & Schacter, D. L. (2014). Imagine all the people: How the brain creates and uses personality models to predict behavior. *Cerebral Cortex, 24*(8), 1979-1987.

Hasson, U., & Malach, R. (2005). *Human brain activation during viewing of dynamic natural scenes.* Paper presented at the Novartis Foundation symposium.

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: The early beginnings. *NeuroImage, 62*(2), 852-855.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science, 293*(5539), 2425-2430.

Heckhausen, H., & Gollwitzer, P. M. (1987). Thought contents and cognitive functioning in motivational versus volitional states of mind. *Motivation and emotion, 11*(2), 101-120.

Heleven, E., & Van Overwalle, F. (2016). The person within: Memory codes for persons and traits using fMRI repetition suppression. *Social Cognitive and Affective Neuroscience, 11*(1), 159-171.

Herrmann, E., Call, J., Hernandez-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science, 317*(5843), 1360-1366.

Holekamp, K. E. (2007). Questioning the social intelligence hypothesis. *Trends in Cognitive Sciences, 11*(2), 65-69.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature, 532*(7600), 453-458.

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron, 76*(6), 1210-1224.

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology, 3*(1), 1-24.

Just, M. A., Cherkassky, V. L., Aryal, S., & Mitchell, T. M. (2010). A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS ONE, 5*(1), e8622.

Kaas, J. H. (1997). Topographic maps are fundamental to sensory processing. *Brain Research Bulletin, 44*(2), 107-112.

Kahneman, D. (2003). Maps of bounded rationality: Psychology for behavioral economics. *American Economic Review, 93*(5), 1449-1475.

Kelley, W. M., Macrae, C. N., Wyland, C. L., Caglar, S., Inati, S., & Heatherton, T. F. (2002). Finding the self? An event-related fMRI study. *Journal of Cognitive Neuroscience, 14*(5), 785-794.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences of the United States of America, 105*(31), 10687-10692.

Kerns, J. G., Cohen, J. D., MacDonald, A. W., Cho, R. Y., Stenger, V. A., & Carter, C. S. (2004). Anterior cingulate conflict monitoring and adjustments in control. *Science, 303*(5660), 1023-1026.

Kestemont, J., Vandekerckhove, M., Ma, N., Van Hoeck, N., & Van Overwalle, F. (2013). Situation and person attributions under spontaneous and intentional instructions: An fMRI study. *Social Cognitive and Affective Neuroscience, 8*(5), 481-493.

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

Kilner, J. M., Friston, K. J., & Frith, C. D. (2007). Predictive coding: An account of the mirror neuron system. *Cognitive Processing, 8*(3), 159-166.

Konkle, T., & Caramazza, A. (2013). Tripartite organization of the ventral stream by animacy and object size. *The Journal of Neuroscience, 33*(25), 10235-10242.

Koster-Hale, J., & Saxe, R. (2013). Theory of mind: A neural prediction problem. *Neuron, 79*(5), 836-848.

Kriegeskorte, N., & Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage, 38*(4), 649-662.

Kriegeskorte, N., Formisano, E., Sorger, B., & Goebel, R. (2007). Individual faces elicit distinct response patterns in human anterior temporal cortex. *Proceedings of the National Academy of Sciences of the United States of America, 104*(51), 20600-20605.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America, 103*(10), 3863-3868.

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

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., . . . Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron, 60*(6), 1126-1141.

Krienen, F. M., Tu, P.-C., & Buckner, R. L. (2010). Clan mentality: Evidence that the medial prefrontal cortex responds to close others. *The Journal of Neuroscience, 30*(41), 13906-13915.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*(4), 978-990.

Leshinskaya, A., Contreras, J. M., Caramazza, A., & Mitchell, J. P. (2017). Neural representations of belief concepts: A representational similarity approach to social semantics. *Cerebral Cortex, 27*(1), 344-357.

Marcus, D. S., Harwell, J., Olsen, T., Hodge, M., Glasser, M. F., Prior, F., . . . Van Essen, D. C. (2011). Informatics and data mining tools and strategies for the human connectome project. *Frontiers in neuroinformatics, 5*, 4.

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

McCormick, T. H., Salganik, M. J., & Zheng, T. (2010). How many people do you know?: Efficiently estimating personal network size. *Journal of the American Statistical Association, 105*(489), 59-70.

McCrae, R. R., & Costa, J., Paul T. (1987). Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology, 52*(1), 81-90.

Mende-Siedlecki, P., Cai, Y., & Todorov, A. (2013). The neural dynamics of updating person impressions. *Social Cognitive and Affective Neuroscience, 8*(6), 623-631.

Mischel, W. (1968). *Personality and assessment*. New York: Wiley.

Mitchell, J. P. (2008). Contributions of functional neuroimaging to the study of social cognition. *Current Directions in Psychological Science, 17*(2), 142-146.

Mitchell, J. P. (2009). Social psychology as a natural kind. *Trends in Cognitive Sciences, 13*(6), 246-251.

Mitchell, J. P., Banaji, M. R., & Macrae, C. N. (2005). The link between social cognition and self-referential thought in the medial prefrontal cortex. *Journal of Cognitive Neuroscience, 17*(8), 1306-1315.

Mitchell, J. P., Cloutier, J., Banaji, M. R., & Macrae, C. N. (2006). Medial prefrontal dissociations during processing of trait diagnostic and nondiagnostic person information. *Social Cognitive and Affective Neuroscience, 1*(1), 49-55.

Mitchell, J. P., Heatherton, T. F., & Macrae, C. N. (2002). Distinct neural systems subserve person and object knowledge. *Proceedings of the National Academy of Sciences of the United States of America, 99*(23), 15238-15243.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2004). Encoding-specific effects of social cognition on the neural correlates of subsequent memory. *The Journal of Neuroscience, 24*(21), 4912-4917.

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron, 50*(4), 655-663.

Mitchell, J. P., Macrae, N. C., & Banaji, M. R. (2005). Forming impressions of people versus inanimate objects: Social-cognitive processing in the medial prefrontal cortex. *NeuroImage, 26*(1), 251-257.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320*(5880), 1191-1195.

Moll, H., & Tomasello, M. (2007). Cooperation and human cognition: The vygotskian intelligence hypothesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 362*(1480), 639-648.

Moran, J. M., Jolly, E., & Mitchell, J. P. (2014). Spontaneous mentalizing predicts the fundamental attribution error. *Journal of Cognitive Neuroscience, 26*(3), 569-576.

Mur, M., Bandettini, P. A., & Kriegeskorte, N. (2009). Revealing representational content with pattern-information fMRI—an introductory guide. *Social Cognitive and Affective Neuroscience, 4*(1), 101-109.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing*. San Francisco, CA: Academic Press.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences, 10*(9), 424-430.

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., & Panksepp, J. (2006). Self-referential processing in our brain--a meta-analysis of imaging studies on the self. *NeuroImage, 31*(1), 440-457.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America, 105*(32), 11087-11092.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The measurement of meaning*. Oxford, England: University of Illinois Press.

Park, S., Konkle, T., & Oliva, A. (2015). Parametric coding of the size and clutter of natural scenes in the human brain. *Cerebral Cortex, 25*(7), 1792-1805.

Parkinson, C., Kleinbaum, A. M., & Wheatley, T. (2017). Spontaneous neural encoding of social network position. *bioRxiv*, 098988.

Parkinson, C., Liu, S., & Wheatley, T. (2014). A common cortical metric for spatial, temporal, and social distance. *The Journal of Neuroscience, 34*(5), 1979-1987.

Peelen, M. V., Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *The Journal of Neuroscience, 30*(30), 10127-10134.

Peirce, J. W. (2007). PsychoPy—psychophysics software in Python. *The Journal of Neuroscience, 162*(1), 8-13.

Posner, J., Russell, J. A., & Peterson, B. S. (2005). The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Development and psychopathology, 17*(03), 715-734.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences of the United States of America, 98*(2), 676-682.

Rauthmann, J. F. (2015). Structuring situational information. *European Psychologist, 20*(3), 176-189.

Reader, S. M., & Laland, K. N. (2002). Social intelligence, innovation, and enhanced brain size in primates. *Proceedings of the National Academy of Sciences of the United States of America, 99*(7), 4436-4441.

Revelle, W., & Rocklin, T. (1979). Very simple structure: An alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research, 14*(4), 403-414.

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *NeuroImage, 22*(4), 1694-1703.

Ross, L. A., & Olson, I. R. (2009). Social cognition and the anterior temporal lobes. *NeuroImage, 49*(4), 3452-3462.

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

Sabatinelli, D., Bradley, M. M., Lang, P. J., Costa, V. D., & Versace, F. (2007). Pleasure rather than salience activates human nucleus accumbens and medial prefrontal cortex. *Journal of Neurophysiology, 98*(3), 1374-1379.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *NeuroImage, 19*(4), 1835-1842.

Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia, 43*(10), 1391-1399.

Schilbach, L., Eickhoff, S. B., Rotarska-Jagiela, A., Fink, G. R., & Vogeley, K. (2008). Minds at rest? Social cognition as the default mode of cognizing and its putative relationship to the "default system" of the brain. *Consciousness and Cognition, 17*(2), 457-467.

Schiller, D., Freeman, J. B., Mitchell, J. P., Uleman, J. S., & Phelps, E. A. (2009). A neural mechanism of first impressions. *Nature Neuroscience, 12*(4), 508-514.

Skerry, A. E., & Saxe, R. (2014). A common neural code for perceived and inferred emotion. *The Journal of Neuroscience, 34*(48), 15997-16008.

Skerry, A. E., & Saxe, R. (2015). Neural representations of emotion are organized around abstract event features. *Current Biology, 25*(15), 1945-1954.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage, 44*(1), 83-98.

Spreng, R. N., Mar, R. A., & Kim, A. S. N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: A quantitative meta-analysis. *Journal of Cognitive Neuroscience, 21*(3), 489-510.

Stolier, R. M., & Freeman, J. B. (2016). Neural pattern similarity reveals the inherent intersection of social categories. *Nature Neuroscience, 19*(6), 795-797.

Szpunar, K. K., Jacques, P. L. S., Robbins, C. A., Wig, G. S., & Schacter, D. L. (2014). Repetition-related reductions in neural activity reveal component processes of mental simulation. *Social Cognitive and Affective Neuroscience, 9*, 712-722.

Tamir, D. I., & Mitchell, J. P. (2010). Neural correlates of anchoring-and-adjustment during mentalizing. *Proceedings of the National Academy of Sciences of the United States of America, 107*(24), 10827-10832.

Tamir, D. I., & Mitchell, J. P. (2011). The default network distinguishes construals of proximal versus distal events. *Journal of Cognitive Neuroscience, 23*(10), 2945-2955.

Tamir, D. I., & Mitchell, J. P. (2013). Anchoring and adjustment during social inferences. *Journal of Experimental Psychology: General, 142*(1), 151-162.

Tamir, D. I., Thornton, M. A., Contreras, J. M., & Mitchell, J. P. (2016). Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences of the United States of America, 113*(1), 194-199.

Tavares, R. M., Mendelsohn, A., Grossman, Y., Williams, C. H., Shapiro, M., Trope, Y., & Schiller, D. (2015). A map for social navigation in the human brain. *Neuron, 87*(1), 231-243.

Todorov, A., Harris, L. T., & Fiske, S. T. (2006). Toward socially inspired social neuroscience. *Brain Research, 1079*(1), 76-85.

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage, 48*(3), 564-584.

van Schaik, C. P., & Burkart, J. M. (2011). Social learning and evolution: The cultural intelligence hypothesis. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 366*(1567), 1008-1016.

Welborn, B., & Lieberman, M. (2015). Person-specific theory of mind in medial pFC. *Journal of Cognitive Neuroscience, 27*(1), 1-12.

Yarkoni, T., Poldrack, R. A., Van Essen, D. C., & Wager, T. D. (2010). Cognitive neuroscience 2.0: Building a cumulative science of human brain function. *Trends in Cognitive Sciences, 14*(11), 489-496.

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., . . . Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by functional connectivity. *Journal of Neurophysiology, 106*(3), 1125-1165.

Zaki, J., López, G., & Mitchell, J. (2014). Activity in ventromedial prefrontal cortex co-varies with revealed social preferences: Evidence for person-invariant value. *Social Cognitive and Affective Neuroscience, 9*(4), 464.

Zaki, J., Schirmer, J., & Mitchell, J. P. (2011). Social influence modulates the neural computation of value. *Psychological Science, 22*(7), 894-900.