Cerebral Cortex, 2017; 1–16

doi: 10.1093/cercor/bhx216
Original Article

ORIGINAL ARTICLE

Theories of Person Perception Predict Patterns of Neural Activity During Mentalizing

Mark A. Thornton and Jason P. Mitchell

Department of Psychology, Harvard University, Cambridge, MA 02138, USA

Address correspondence to Mark A. Thornton, Department of Psychology, Harvard University, 33 Kirkland Street, Cambridge, MA 02138, USA.
Email: mthornton@fas.harvard.edu.

Abstract

Social life requires making inferences about other people. What information do perceivers spontaneously draw upon to make such inferences? Here, we test 4 major theories of person perception, and 1 synthetic theory that combines their features, to determine whether the dimensions of such theories can serve as bases for describing patterns of neural activity during mentalizing. While undergoing functional magnetic resonance imaging, participants made social judgments about well-known public figures. Patterns of brain activity were then predicted using feature encoding models that represented target people's positions on theoretical dimensions such as warmth and competence. All 5 theories of person perception proved highly accurate at reconstructing activity patterns, indicating that each could describe the informational basis of mentalizing. Cross-validation indicated that the theories robustly generalized across both targets and participants. The synthetic theory consistently attained the best performance—approximately two-thirds of noise ceiling accuracy—indicating that, in combination, the theories considered here can account for much of the neural representation of other people. Moreover, encoding models trained on the present data could reconstruct patterns of activity associated with mental state representations in independent data, suggesting the use of a common neural code to represent others' traits and states.

Key words: functional magnetic resonance imaging, multivoxel pattern analysis, social cognition, theory of mind

Introduction

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. Here, we test 4 specific instances of this dimensional meta-theory of person perception, as well as one synthetic theory derived from their components. We aim to examine whether any of these theories can explain the organization of person knowledge, measured using pattern analysis of brain activity, and if so, which perform better or worse than others, and how well they all compare to a hypothetical ideal theory.

Theories of the class we consider 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 2 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, 4 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.

Like the stereotype content model, the 3 other theories we consider were also developed to address relatively specific phenomena. The 5-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 (McCrae and Costa 1987; Goldberg 1990). 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 2-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 2-factor theory of social face perception (Oosterhof and 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 2 dimensions: trustworthiness and social dominance. Certain facial features are associated with each dimension: for instance, wide set eyes and upturned mouths drive trustworthiness judgments.

In addition to these 4 extant theories, we also examine a fifth, synthetic theory that combines 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.

Despite the success of the 4 extant 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 contexts. Even if the meta-theory of a representational space for other people is true in general, it is not clear which specific dimensions embody the successful instantiation of the 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 accounts.

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 cognitive neuroscience (Mitchell et al. 2008; Huth et al. 2016), into the social domain. This framework combines functional magnetic resonance imaging (fMRI) and multivoxel pattern analysis (MVPA). This approach has multiple advantages, several due to the nature of fMRI as a measure. First, BOLD fMRI is an implicit measure, as people are not consciously aware of oxygenated blood flow in their brains, and this 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 such responses need not be the primary source of outcome data in an fMRI paradigm.

The modeling 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 (Mitchell et al. 2002; Saxe and Kanwisher 2003; Saxe and Wexler 2005; Mitchell 2009). 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 and Bandettini 2007; Mur et al. 2009). This approach has already borne fruit in social neuroscience, yielding insight into the mental and neural organization of mental state representations (Skerry and Saxe 2015; Tamir et al. 2016) and social categories (Stolier and 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 2 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" modeling. In this approach, one or more theories are compared in terms of their ability to predict patterns of brain activity associated with particular stimuli (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.

We supplement feature space modeling with the use of representational similarity analysis (RSA), which we use to assess the ability of each individual dimension to explain the (dis)similarity between target person-specific patterns of brain activity (Kriegeskorte et al. 2008). We also use this approach to test 2 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 RSA 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 modeling, we can directly test 5 theories of person perception against the null of baseline performance and against each other. 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. Direction comparisons of the theories would reveal which provide better accounts for the neural organization of person knowledge. 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.

Materials and Methods

Code and Data Availability

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 (OSF) (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 conducted on the same fMRI scanner. 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 pretested 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 (GLM), and pattern dissimilarity (correlation distance) between these patterns was calculated within regions showing a univariate effect of mental state identity ($P < 0.0001$, 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 $r = 0.05$, 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. RSA was conducted on each participant in the bootstrapped sample, regressing neural pattern similarity between the 60 mental states onto the behavioral predictions of the 3 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 0.95.

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 2 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 2 properties, we employed a 2-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 web scraping 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, nonword artifacts) we required that words included in the master list appear at least twice each in at least 2 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.

Pretesting

After the completion of the Internet-based stimulus selection, 2 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% 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 3 more potentially problematic targets: Jeffrey Dahmer, due to the extreme ratings, Steve McQueen, due to the presence 2 relatively famous people sharing this name, and Anne Frank, due to being the only nonadult 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 1). Each participant was asked to rate all 60 targets with respect to only 1 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 13 dimensions in total: warmth, competence, agency, experience, trustworthiness, dominance, openness to experience, conscientiousness, extraversion, agreeableness, neuroticism, attractiveness, and intelligence. Although the last 2 of these dimensions do not together constitute an extant model, we included them because they are extremely

Table 1 The reliability of norming ratings of targets on psychological dimensions

Dimension	Theory	n	Mean inter-rater r	Cronbach's α
Agency	Mind perception	60	0.13	0.90
Agreeableness	Five factor personality	58	0.31	0.96
Attractiveness	None	71	0.29	0.97
Competence	Stereotype content	69	0.38	0.98
Conscientiousness	Five factor personality	65	0.41	0.98
Dominance	Social face perception	72	0.32	0.97
Experience	Five factor personality	52	0.35	0.97
Extraversion	Five factor personality	72	0.40	0.98
Intelligence	None	65	0.52	0.99
Neuroticism	Five factor personality	62	0.26	0.96
Openness	Five factor personality	66	0.16	0.93
Trustworthiness	Social face perception	64	0.39	0.98
Warmth	Stereotype content	65	0.30	0.97

widely discussed features of other people in both the scientific literature and lay parlance. Ratings for each target were averaged across participants to provide a 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.

Dimensionality Reduction

To generate an optimal synthetic theory, principal components analysis (PCA) was applied to the correlation matrix between the 13 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 3 principal components. These components loaded mostly highly onto dominance, warmth, and extraversion (reflected), respectively (Fig. 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 0.88 and a minimum of 0.71. Component 1 correlated with component 2 at $r = 0.26$, and with component 3 at $r = 0.36$ components 2 and 3 were correlated at $r = 0.22$. While nontrivial, these correlations were sufficiently small to minimally complicate interpretation, relative to an orthogonal model. The scores from the 3 factors were used in place of ratings for the purposes of the constructing encoding models for the resulting 3-component synthetic theory. Note that the component scores remain virtually identical ($rs > 0.99$) if the same PCA is applied to theoretical dimensions only, excluding the additional dimensions of intelligence and attractiveness.

Experimental Paradigm

The imaging paradigm consisted of a modified version of a mentalizing task used in previous research (Mitchell et al. 2006; Tamir and 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 3.25 s response window. This period would be followed by a 250 ms minimum fixation period, and a variable jitter fixation period (mean jitter = 1.33 s, approximately Poisson distributed in 2 s increments). Each run of the experiment consisted of 60 trials: one for each of the target people. The runs were also balanced with respect

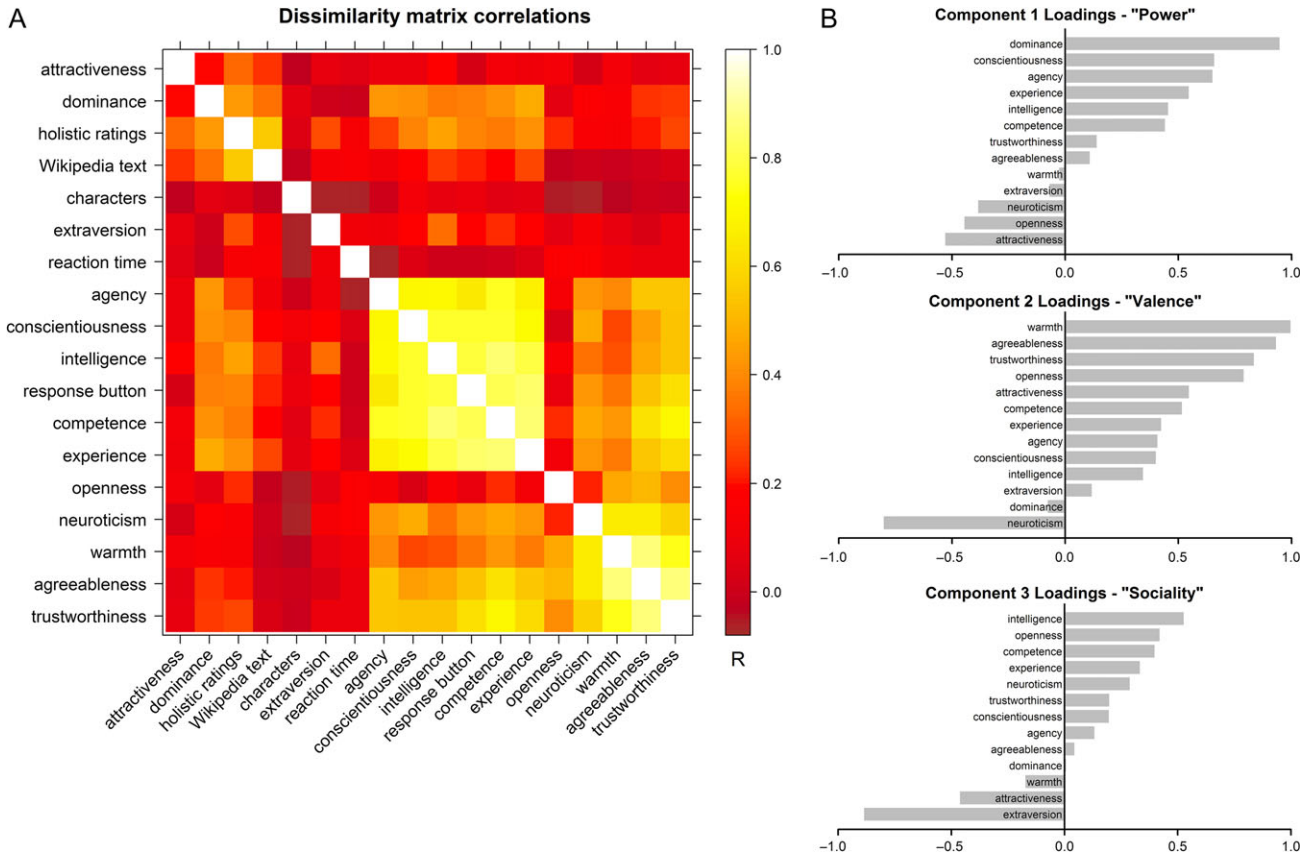

Figure 1. Dimensions organizing target people. (*A*) Correlations between dissimilarity matrices produced by taking the absolute differences in positions of targets on each of the 13 rated dimensions. Two pairwise measures of (dis)similarity were also included: textual dissimilarity of Wikipedia biography 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.

to the social judgment items, with 5 of each of the 12 appearing in every run. An additional 6 s 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 2 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 3 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 2) used in the imaging paradigm were selected for minimal redundancy from a larger set of 24 items tested in 2 pilot versions of the behavioral task conducted outside the scanner. The initial set of 24 pilot items was selected manually based on 3 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 2, 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.

One potential concern regarding the imaging results might be that the social judgment items chosen for the study could be biased toward 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 RSA comparing the 4 extant theories to the social judgment items. For each of the 4 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, $r = 0.69$; social face perception, $r = 0.65$; 5 factor personality, $r = 0.71$; mind perception, $r = 0.76$. 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 prescan 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 modeling, 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) 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 2 mm thickness, TR = 2000 ms, TE = 30 ms, flip angle = 80°, in-plane resolution = 2.00 × 2.00 mm, matrix size = 108 × 108 voxels, 162 measurements per run). Functional images were preprocessed and analyzed using SPM8 (Wellcome Department of Cognitive Neurology) 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

Table 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 traveling 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

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 modeled 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, 6 motion realignment parameters, and outlier time points (defined by the Artifact Detection Toolbox). Regression coefficient maps from this analysis were smoothed with a 6 mm FWHM Gaussian kernel to improve interparticipant 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. However, the representational similarity analyses we conducted could not be influenced by 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

Interparticipant 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 (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 0.01) 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 half of the data set. 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 toward 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 Modeling

Feature space modeling is a form of generative computational encoding modeling of neural activity (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 4 theories of person perception discussed above, and the synthetic 3-component theory generated by PCA, were analyzed and compared through feature space modeling. 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 (CIs) around mean performance were obtained via percentile bootstrapping of participants (with 10 000 samples). Percentile bootstrapping was again used to test pairwise difference in model performance via CIs around mean differences with 100 000 bootstrap samples. The resulting CIs were Bonferroni-corrected to account for all pairwise comparisons, and thus statistically significant results pass an uncorrected 0.005 threshold.

Direct null hypothesis significance testing of model performance (versus chance) was also carried out non-parametrically via permutation-based prevalence testing (Allefeld et al. 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 1 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% CI 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., $0.05^{(1/29)}$) 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 interparticipant 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 data set. 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 3-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 3 factors extracted in the present study: power, sociality, and valence respectively. Using the 2 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 data set.

Note that this validation approach represents a particularly stringent challenge for the feature encoding models due to the differences between the 2 data sets. 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 2 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 (Kriegeskorte et al. 2008) 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 2 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, 2 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 pretesting, and the similarity measure computed from the targets' Wikipedia text. These measures were converted to dissimilarities by reflection where necessary. We also included 3 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 nonredundant lower-triangular elements of each square dissimilarity matrix. Significance testing was conducted directly via permutation testing ($N = 1000$) 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 5 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 et al. 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 2 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 gray 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 solely from the visual feedback. A plausible alternative might be that responses only correlate with pattern similarity because responses are themselves primarily influenced by social dimensions.

Permutation testing and bootstrapping were repeated for each variant. Additionally, noise ceilings were calculated for each of the 4 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 4 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 2 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. Responses from the imaging task were combined with participants prescanning 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 modeling, 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. Analysis of the behavioral results replicated a number of well-established findings in the literature. Increased similarity between the participant and the target predicted diminished self-other discrepancy ($b = -0.04$, $\beta = -0.10$, $P < 2 \times 10^{-16}$), as predicted by the similarity-contingency model (Ames 2004). Longer reaction time predicted greater self-other discrepancy ($b = 0.12$, $\beta = 0.07$, $P < 3 \times 10^{-4}$), replicating findings on egocentric anchoring and adjustment in mentalizing (Epley et al. 2004). Finally, consistent with recent results (Tamir and 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 ($b = 0.03$, $\beta = 0.04$, $P < 2 \times 10^{-5}$). 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 4 extant theories were established through additional online norming surveys. The idiosyncrasy of such perceptions was reflected in modest average inter-rater correlations (Table 1). However, the composites all ultimately achieved high levels of reliability, with an average Cronbach's $\alpha = 0.96$ (SD = 0.02). We observed large correlations between the dimensions of person perception theories, indicating considerable redundancy between the predictions of the theory dimensions (Fig. 1A). We conducted PCA on the traits to synthesize a single parsimonious theory that encapsulates the unique predictions of all rated dimensions. The 3 components of the optimal solution loaded most heavily on dominance, warmth, extraversion (Fig. 1B). To distinguish between manifest and latent variables, we labeled 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 (Mitchell et al. 2002; Saxe and Kanwisher 2003; Saxe and Wexler 2005; Mitchell 2009), including medial prefrontal cortex, posterior parietal cortex, the temporoparietal junction, and portions of the lateral temporal lobe (Fig. 2). This confirms that thinking about different well-known target people elicits reliability differentiable neural activity within the social brain network.

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 = 0.54). The lowest pattern reliability was observed in somatosensory (α = 0.13) and visual cortices (α = 0.20). Intermediate values were observed for other association cortex networks: dorsal attention—0.29, ventral attention—0.21, limbic—0.24, and frontoparietal—0.47. The residual pattern similarity reliability in the default network—controlling for pattern similarity in the visual and somatosensory cortices—was 0.49. 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 any possibility of visual or motor confounds might account for observed effects.

Feature Space Modeling

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 < 0.001) performance for all 5 theories (Fig. 3A) as assessed bootstrapping. The lower bound of the 95% CI 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 0.070 for the 5-factor model (Cohen's d = 1.89), 0.080 for the mind perception model (d = 2.16), 0.083 for the social face perception model (d = 1.94), 0.081 for the stereotype content model (d = 1.90), and 0.104 for synthetic the 3 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). The 3-component synthetic model achieved the highest performance in all brain network 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.

Pairwise indirect difference tests (via bootstrapping) in the reliability selected regions indicated that the 3-component model significantly outperformed the 4 extant models, and the social face perception model significantly outperformed the 5 factor personality model (Table 4). Note that the latter outcome demonstrates that more parsimonious theories (the 2-D face perception model) can indeed outperform more complex theories (the 5-factor model) in the present framework, despite the absence of discounting for higher dimensional theories.

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 5 models performed significantly above chance (Ps < 0.001). 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: 0.063 for the 5-factor model, 0.053 for the mind perception model, 0.056 for the social face perception model, 0.061 for the stereotype content model, and 0.072 for the 3 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 ds 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 5-factor model—which had the worst performance in the leave-one-target-out case—here outperformed the other 3 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 5 models achieved approximately half to two-thirds of the maximum performance possible given the variance shared across participants: 0.59 for the 5-factor model, 0.49 for the mind perception model, 0.52 for the social face perception model, 0.56 for the stereotype content model, and 0.67 for the 3 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 (Fig. 4) within each voxel of the feature-selected region.

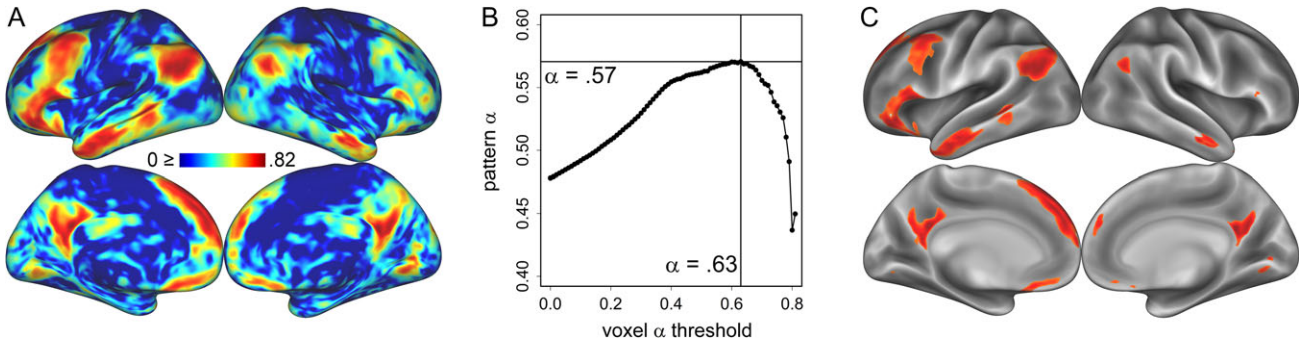

Figure 2. Reliability-based feature selection. Univariate reliability across targets (A) was calculated for each voxel via the formula for standardized Cronbach's α. The reliability of the correlation matrix between patterns was assessed at each voxelwise reliability threshold between 0 and the observed maximum (B). The voxelwise reliability threshold that maximized pattern correlation reliability was used to select voxels for further analysis (C).

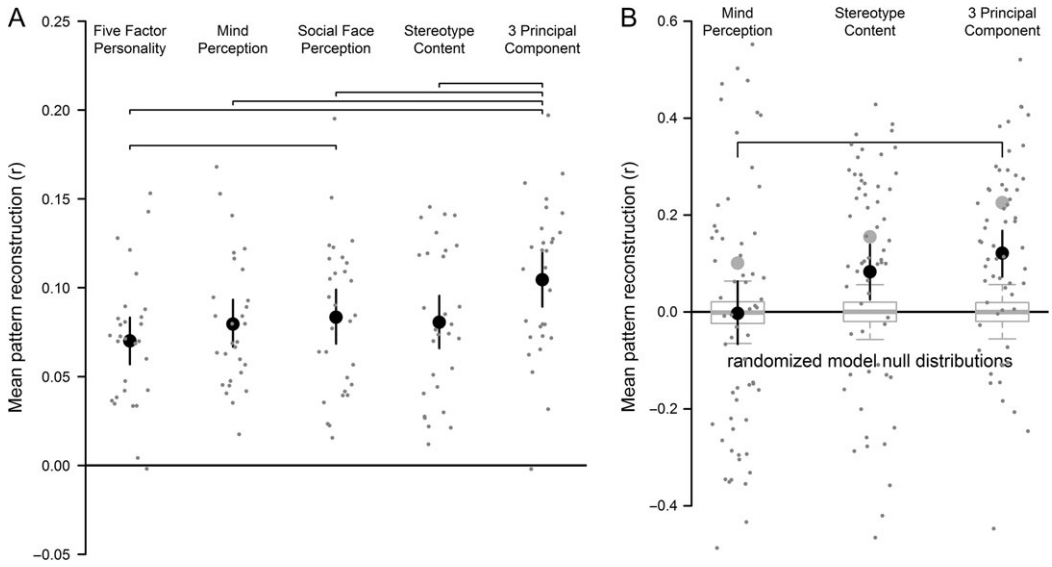

Figure 3. Feature space modeling performance. Large black circles indicate mean pattern reconstruction, and the error bars around them indicate 95% CIs. Brackets indicate significant pairwise differences. Boxplots indicate empirical null distributions with whiskers at the 0.975 and 0.025 quantiles. Model performance in leave-one-target-out cross-validation analysis (*A*) was above chance for all 5 models. Small gray points indicate the performance of individual participants. Across-data set pattern reconstruction accuracy (*B*) was above chance for 2 of 3 theories—stereotype content and the synthetic 3-component model. Small gray circles indicate pattern reconstruction accuracy for each of 60 states.

Table 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	Three 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

Table 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	−0.010	−0.027	0.009
2	Five factor personality	Social face perception	−0.013	−0.027	0.000
3	Five factor personality	Stereotype content	−0.010	−0.022	0.001
4	Five factor personality	3-component	−0.034	−0.044	−0.024
5	Mind perception	Social face perception	−0.004	−0.018	0.012
6	Mind perception	Stereotype content	−0.001	−0.018	0.014
7	Mind perception	3-component	−0.025	−0.041	−0.010
8	Social face perception	Stereotype content	0.003	−0.011	0.018
9	Social face perception	3-component	−0.021	−0.033	−0.009
10	Stereotype content	3-component	−0.024	−0.035	−0.012

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

In a third validation method, we used a completely independent data set 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, 2 of the 3 theories successfully reconstructed patterns in the other data set (Fig. 3*B*)—the stereotype content model, with a mean pattern reconstruction of 0.083 ($P < 0.005$), and the 3-component model, with a mean pattern reconstruction of 0.12 ($P < 0.001$). The mind perception model failed to significantly predict, with a mean pattern accuracy of −0.0003 ($P > 0.1$). Both successful trait-trained theories achieved 0.54 of their models possible performance by this metric. Pairwise difference testing indicated that the synthetic 3-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 ($rs < 0.1$)

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 3-component model (Fig. 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 $r = 0.43$ for power/rationality, $r = 0.28$ for valence, and $r = 0.29$ for sociality/social impact. This suggests that the dimension of power/rationality is the most conserved across the trait-state divide.

Representational Similarity Analysis

RSA was conducted to test the ability of individual dimensions to explain the (dis)similarity of patterns of neural activity associated with different target people (Fig. 6). The 2 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, many of the predictors

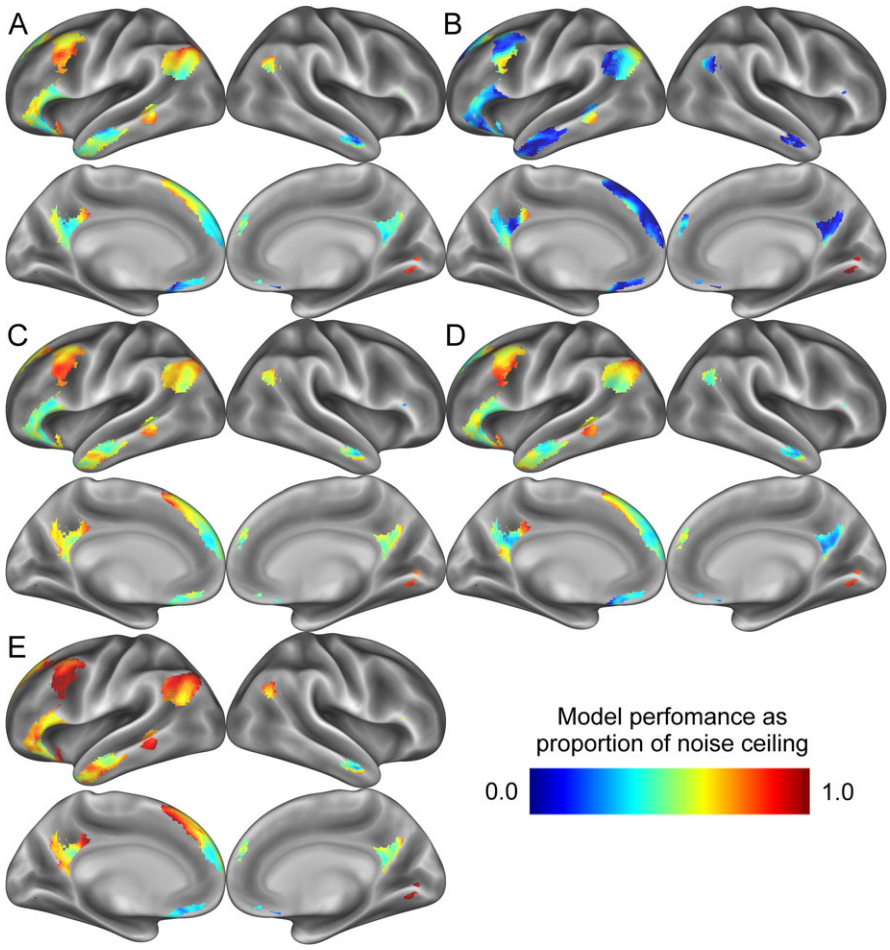

Figure 4. Voxelwise model performance in leave-one-participant-out cross-validation. Voxelwise correlations between model predictions and actual activity in a leave-one-participant-out cross-validation analysis are shown for each of the five theories under consideration: the 5-factor model (*A*), the mind perception model (*B*), the social face perception model (*C*), the stereotype content model (*D*), and the synthetic 3-component model (*E*).

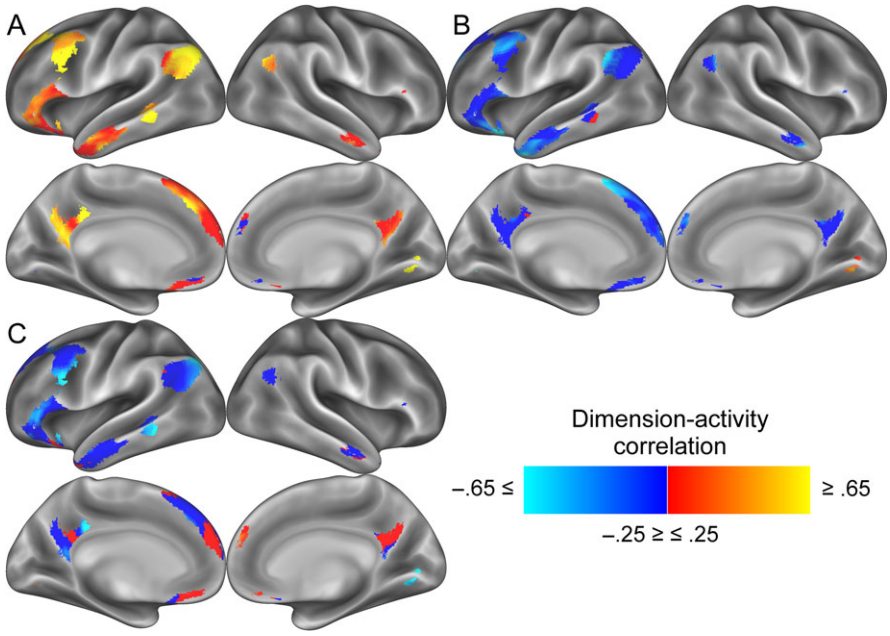

Figure 5. The functional topology of the 3-component encoding model. Colors reflect a voxelwise mapping of correlations between average univariate activity and the 3 principal components of the synthetic theory: power (A), valence (B), and sociality (C). Activity in orange areas is thus positively associated with thinking about dominant, warm, and extraverted public figures, respectively.

that 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.

Multiple regression RSA indicated considerable redundancies among the 4 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 0.61 for the 5-factor model, 0.40 for the mind perception model, 0.83 for the social face perception model, and 0.45 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 0.36 to 0.69. The predictions of the mind perception model were the most unique, with correlations of only $r = 0.35$ with the 5-factor model, and rs = 0.36 for both the social face perception and stereotype content model. The 5-factor model was involved in the 2 highest correlations between model fitted values: $r = 0.69$ with the social face perception model, and $r = 0.63$ with the stereotype content model. The social face perception and stereotype content models made moderately similar accurate predictions about neural pattern similarity: $r = 0.47$. 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 4 prominent theories of person perception (McCrae and Costa 1987; Goldberg 1990; Fiske et al. 2002; Gray et al. 2007; Cuddy et al. 2008; Oosterhof and Todorov 2008)—and 1 synthetic theory combining their dimensions—predict patterns of neural activity elicited by mentalizing. The primary result was clear: encoding models based on all 5 of theories performed substantially above chance. Indeed, each encoding model positively predicted the brain activity of nearly every participant, and performed between half and two-thirds as well as a hypothetical ideal theory. This outcome strongly supports the idea—shared across the tested theories—that people represent others within a multidimensional social representational space. This indicates that a distributed population code linearly encodes psychological dimensions used to organize social knowledge. Moreover, the fact that the theory-predicted neural activity patterns were elicited by making a variety of 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 3 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 5 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 2 theories of person perception—the stereotype content model and our synthetic 3-component model—to reconstruct activity patterns across-data sets also carries significantly implications. This generalization indicates that the

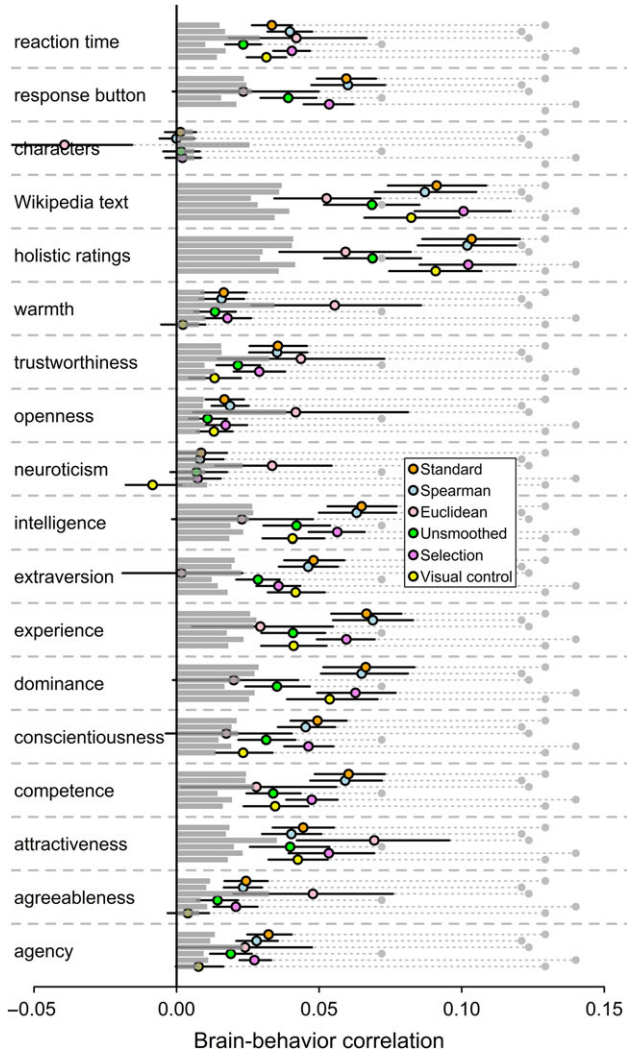

Figure 6. RSA. The correlation between neural pattern dissimilarity and 15 different predictors was assessed under a standard analysis reflecting the feature space encoding models and 5 reasonable methodological variants. Colored circles indicate the magnitude of the correlations for each predictor, with 95% CIs around them. Gray circles indicate noise ceilings, and gray bars indicate the 0.975 quantile of permutation distributions.

neural patterns that encode theoretical dimensions are robust to many low-level differences between data sets, 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 data sets without the need to train the models separately for each sample or participant.

Moreover, the encoding models generalized 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 psychologists (Mischel 1968).

Comparing the 4 extant theories of person perception revealed surprisingly similar performance. There were few pairwise differences in model performance, and most that were observed stemmed solely from the superior performance of the synthetic theory. Moreover, all of the theories appeared to explain neural activity to similar degrees across different brain regions. The behavioral predictions of the theoretical dimensions were often highly correlated, and the accurate neural predictions of the extant theories also overlapped considerably. These results suggest that the study of several different phenomena—intergroup affect, personality, mind perception, and face perception—may have given rise to generally similar theories of person perception. This may suggest that the brain applies a single unifying approach to organizing social knowledge, rather than developing specialized representational frameworks for different social domains.

Across all testing approaches, our 3-component synthesis of the dimensions of the extant theories outperformed all of the original theories. As stated above, it also closely approximates—and cross-decodes—a similar model of mental state representation. On this basis, we suggest that people may consistently consider 3 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). These qualities will depend on both the traits and states of others, accounting for the partially shared representational space. Determining others' coordinates within such a three-dimensional space would make it possible to cogently engage in a wide range of adaptive behavioral responses. Directly testing this adaptive account of social knowledge organization should be a topic for future research. If validated, this framework might suggest additional dimensions which help complete our theories of person perception.

The performance of the synthetic principal component encoding model could be compared to noise ceilings in 2 of the 3 cross-validation procedures we conducted (across participants and data sets). 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 4 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 neural 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 and 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 and Tenenbaum 2008) to condition-rich data sets 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. Testing under this assumption allows us to go beyond merely asserting that the social brain network contains person-specific activity patterns, to describe the nature of the encoding scheme within the network—that is, the meta-theory of a representational space for other people, as a opposed to other schemes such as sparse or temporal coding. The linear dimensional meta-theory is among the simplest imaginable population coding schemes, and yet despite this naïve assumption, we observed remarkably robust performance. It is possible that modeling based on more complex assumptions—for example, genuinely bipolar dimensions, or nonlinear 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 results suggest 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. Thus, these results validate the overall meta-theory of a representational space for people, as well as the specific instantiations of that meta-theory which we borrowed from the psychological literature.

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). Purely behavioral results from our dimension pretesting studies also support recent research that suggests that valence can be separated into social and moral components (Brambilla and 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. The 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 3 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. Because the purely visual character length confound had no effect, 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. Moreover, visual cortex activity cannot explain the vast majority of pattern similarity within the social brain network, nor can visual confounds explain any of the cross-data set accuracy of the encoding models. Ultimately more research is needed to decouple the social properties of target people from low-level task features. Adopting more naturalistic approaches may also help address this issue and others by limiting low-level confounds to those found ecologically in typical experience.

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 assemble 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.

Funding

National Science Foundation Graduate Research Fellowship (DGE1144152) and by The Sackler Scholar Program in Psychobiology (to M.A.T.).

Notes

The authors would like to thank Juan Manuel Contreras, Spencer Dunleavy, Talia Konkle, Abigail Orlando, Joakim Norberg, Franchesca Ramirez, Ryan Song, and Diana Tamir for their assistance. *Conflict of Interest*: None declared.

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 DR. 2004. Strategies for social inference: a similarity contingency model of projection and stereotyping in attribute prevalence estimates. J Pers Soc Psychol. 87:573–585.

Brambilla M, Leach CW. 2014. On the importance of being moral: the distinctive role of morality in social judgment. Soc Cogn. 32:397–408.

Cuddy AJ, Fiske ST, Glick P. 2008. Warmth and competence as universal dimensions of social perception: the stereotype content model and the BIAS map. Adv Exp Soc Psychol. 40: 61–149.

Davis T, LaRocque KF, Mumford JA, Norman KA, Wagner AD, Poldrack RA. 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.

Epley N, Keysar B, Van Boven L, Gilovich T. 2004. Perspective taking as egocentric anchoring and adjustment. J Pers Soc Psychol. 87:327–339.

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. J Pers Soc Psychol. 82:878–902.

Goldberg LR. 1990. An alternative "description of personality": the big-five factor structure. J Pers Soc Psychol. 59: 1216–1229.

Goodwin GP. 2015. Moral character in person perception. Curr Dir Psychol Sci. 24:38–44.

Gray HM, Gray K, Wegner DM. 2007. Dimensions of mind perception. Science. 315:619.

Gross CG. 2002. Genealogy of the "grandmother cell". Neuroscientist. 8:512–518.

Hassabis D, Spreng RN, Rusu AA, Robbins CA, Mar RA, Schacter DL. 2014. Imagine all the people: How the brain creates and uses personality models to predict behavior. Cereb Cortex. 24:1979–1987.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science. 293: 2425–2430.

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature. 532:453–458.

Kaas JH. 1997. Topographic maps are fundamental to sensory processing. Brain Res Bull. 44:107–112.

Kemp C, Tenenbaum JB. 2008. The discovery of structural form. Proc Natl Acad Sci USA. 105:10687–10692.

Khaligh-Razavi S-M, Kriegeskorte N. 2014. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol. 10:e1003915.

Kriegeskorte N, Bandettini P. 2007. Analyzing for information, not activation, to exploit high-resolution fMRI. Neuroimage. 38:649–662.

Kriegeskorte N, Mur M, Bandettini PA. 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci. 2:4.

McCrae RR, Costa PT Jr. 1987. Validation of the five-factor model of personality across instruments and observers. J Pers Soc Psychol. 52:81–90.

Mischel W. 1968. Personality and assessment. New York: Wiley.

Mitchell JP. 2009. Social psychology as a natural kind. Trends Cogn Sci. 13:246–251.

Mitchell JP, Heatherton TF, Macrae CN. 2002. Distinct neural systems subserve person and object knowledge. Proc Natl Acad Sci USA. 99:15238–15243.

Mitchell JP, Macrae CN, Banaji MR. 2006. Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. Neuron. 50:655–663.

Mitchell TM, Shinkareva SV, Carlson A, Chang K-M, Malave VL, Mason RA, Just MA. 2008. Predicting human brain activity associated with the meanings of nouns. Science. 320:1191–1195.

Mur M, Bandettini PA, Kriegeskorte N. 2009. Revealing representational content with pattern-information fMRI—an introductory guide. Soc Cogn Affect Neurosci. 4:101–109.

Newell A. 1973. You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In: Chase WG, editor. Visual information processing. San Francisco, CA: Academic Press, p. 1–24.

Oosterhof NN, Todorov A. 2008. The functional basis of face evaluation. Proc Natl Acad Sci USA. 105:11087–11092.

Saxe R, Kanwisher N. 2003. People thinking about thinking people: the role of the temporo-parietal junction in "theory of mind". Neuroimage. 19:1835–1842.

Saxe R, Wexler A. 2005. Making sense of another mind: the role of the right temporo-parietal junction. Neuropsychologia. 43:1391–1399.

Skerry AE, Saxe R. 2015. Neural representations of emotion are organized around abstract event features. Curr Biol. 25:1945–1954.

Stolier RM, Freeman JB. 2016. Neural pattern similarity reveals the inherent intersection of social categories. Nat Neurosci. 19:795–797.

Tamir DI, Mitchell JP. 2010. Neural correlates of anchoring-and-adjustment during mentalizing. Proc Natl Acad Sci USA. 107: 10827–10832.

Tamir DI, Mitchell JP. 2013. Anchoring and adjustment during social inferences. J Exp Psychol Gen. 142:151–162.

Tamir DI, Thornton MA, Contreras JM, Mitchell JP. 2016. Neural evidence that three dimensions organize mental state representation: rationality, social impact, and valence. Proc Natl Acad Sci USA. 113:194–199.

Tavares RM, Mendelsohn A, Grossman Y, Williams CH, Shapiro M, Trope Y, Schiller D. 2015. A map for social navigation in the human brain. Neuron. 87:231–243.

Yeo BTT, Krienen FM, Sepulcre J, Sabuncu MR, Lashkari D, Hollinshead M, Roffman JL, Smoller JW, Zollei L, Polimeni JR, et al. 2011. The organization of the human cerebral cortex estimated by functional connectivity. J Neurophysiol. 106: 1125–1165.