COMMENTARY

How the brain represents other minds

Julien Dubois[a,1] and Ralph Adolphs[a,1]

How does the brain represent the world? Sensory neuroscience has given us a detailed window into how the brain represents physical objects in our environment: For instance, the shape, color, and direction of motion of visual stimuli are represented in an orderly fashion in higher order visual cortices. But we also represent social objects: other people and their thoughts, beliefs, and feelings. How is that kind of knowledge represented in the brain? In an ambitious new study in PNAS, Tamir et al. (1) used functional MRI (fMRI) to argue that our brains represent other minds along three broad dimensions: social impact, rationality, and valence.

The approach taken by the authors is straightforward to understand by analogy. For example, we could represent any specific place on earth we have visited as a unique combination of a large number of variables: altitude, temperature, humidity, light, chemical composition of the air, and so forth. Despite the enormous number of different variables, you could quickly tell me which of these visits were more similar to one another: Standing on top of Mont Blanc would be more similar to standing on top of Mount Whitney than to lying on your towel on a beach in Aruba. To judge such similarities, you do not need to recollect all of the details; a few coarse dimensions suffice. We do the same thing when we think about other people, Tamir et al. (1) argue. Psychologists have attempted to capture the specific dimensions by which we represent others in several theories, from which the study extracted 16 dimensions for further investigation (Fig. 1*A*). Importantly, prior work on these dimensions was based largely on theory and on behavioral data. Tamir et al. (1) looked to the brain for further evidence.

Human participants underwent fMRI while they imagined another person experiencing a number of different mental states (60 mental states; table S1 in ref. 1). On a given trial, a word, such as "awe," was shown together with two phrases, such as "seeing the Pyramids" or "watching a meteor shower," and participants had to decide which phrase best fit the word. As in the example, there was no right or wrong answer, effectively ensuring that the participants' brains were hard at work thinking about the mental state under scrutiny. First picking the 60 mental state terms and the phrases, however, required some work.

The study is a tour de force in design. An initial list of 166 mental states (i.e., 166 "locations" in the space whereby we represent other people's minds) was categorized into specific dimensions by a separate group of subjects (through Amazon's Mechanical Turk over the Internet). This list was whittled down to an optimal subset by maximizing the distinctiveness of each state (i.e., highest loading on a subset of 16 dimensions and lowest loading on the rest) while minimizing the similarity between selected states. For each of the final 60 states thus produced, 36 concise and believable scenarios were generated and once more validated with an independent set of participants over the Internet, who produced ratings (about 65 for each scenario) of the association of each scenario with the intended mental state. The authors wanted to use 16 scenarios per mental state for their fMRI study (as suggested by a power analysis; discussed below), which not only had to be most representative of their intended mental state, but all of the different mental states also had to be captured equally well by their scenarios, and even simple properties like word counts needed to be similar across scenarios. A genetic algorithm jointly optimized all these objectives. A power analysis suggested a design with 20 subjects each scanned for about 2 h (16 runs with 60 trials each, corresponding to the 60 mental state terms). All in all, this study is a poster child for how one should design an fMRI experiment.

After collecting all these data, the goal was now to examine the mental state space inferred from the neural data (the brain's dimensions for representing other people's mental states) and to compare this neural space with the dimensions suggested by psychological theories and behavioral data. A recently developed analytical technique was perfectly suited for this challenge: representational similarity analysis (RSA), which was initially introduced in visual neuroscience (2) but has now been applied to many fields, including social neuroscience (3). On the one hand,

[a]Division of Humanities and Social Sciences, California Institute of Technology, Pasadena, CA 91125

Author contributions: J.D. and R.A. wrote the paper.

The authors declare no conflict of interest.

See companion article on page 194.

[1]To whom correspondence may be addressed. Email: radolphs@caltech.edu or jcrdubois@gmail.com.

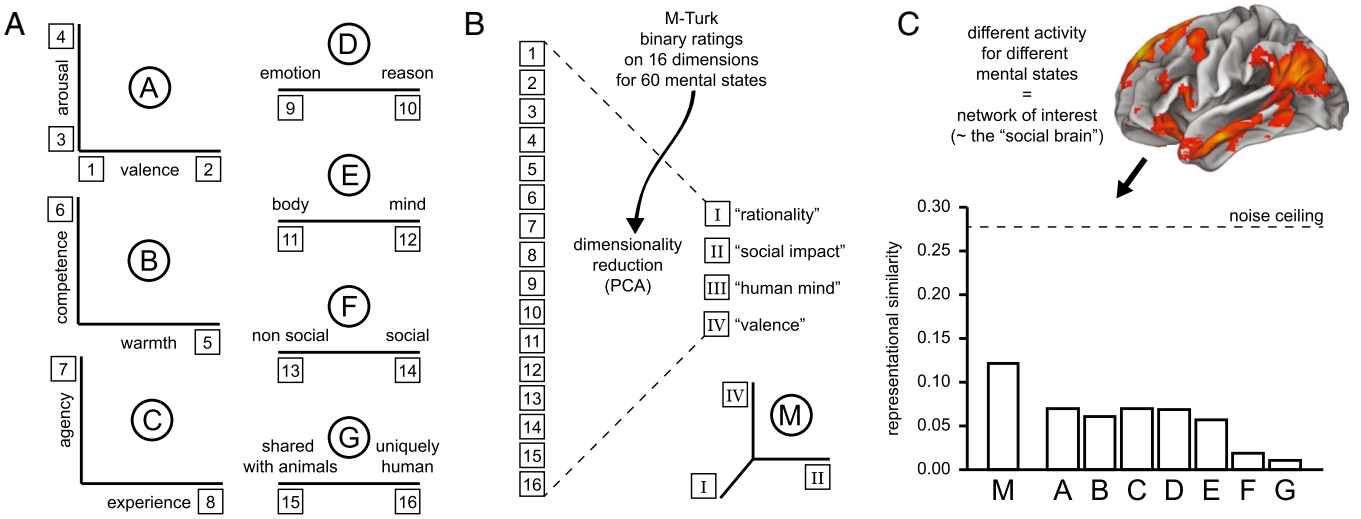

Fig. 1. Summary of the study by Tamir et al. (1). (*A*) Seven existing psychological models (A through G) encompassed different dimensions (1 through 16) along which the mental states of others are represented psychologically. (*B*) Online ratings for 60 mental states on these 16 dimensions, together with dimensionality reduction, resulted in four new dimensions for mental state representation. Three of these dimensions fit the data best, and were combined into a new 3D model, M. PCA, principal components analysis. (*C*) Brain regions differentially activated by mental states (statistical parametric map displayed on brain surface) were used in a representational similarity analysis to test extant psychological models (A through G) as well as the new model M. Model M was found to fit the neural data best, although there is room for improvement because it was still below the noise ceiling [the best possible fit (dotted line)].

patterns of brain activity were extracted from the fMRI data: The authors ran a classical general linear model and made a separate map (*t*-values, across 16 runs) of the whole-brain blood oxygen level-dependent (BOLD) activity triggered by each of the 60 mental states. The similarity of these patterns was computed using a Pearson correlation coefficient, resulting in a 60 × 60 similarity matrix. On the other hand, the psychological similarity of the conditions of interest was assessed relative to seven psychological models (A through G in Fig. 1*A*) onto which the mental states could be mapped (using the ratings that were initially acquired over the Internet). A psychological similarity matrix (again, of dimensions 60 × 60) was built for each candidate psychological model. Finally, the neural similarity matrix from the fMRI data was compared with each of the psychological model similarity matrices in turn, using linear regression or Pearson correlation. Which similarity structure derived from the psychological models showed the best correspondence to the neural similarity structure derived from the fMRI data?

The authors ran the RSA in two ways: first, using an anatomical "searchlight" approach whereby a sphere (with a radius of about 9 mm) was moved throughout the entire brain to discover any local representations; second, using a network-of-interest approach, which consisted of selecting those brain regions differentially activated by thinking about mental states in the experiment [overlapping with the known "social brain" (4, 5)]. Restricting ourselves here to the results from the network-of-interest analysis, the similarity structure derived from the fMRI data fit equally well for the first five existing psychological models (A through E, $r \sim 0.2$), but not significantly for the last two (F and G in Fig. 1*C*).

To go beyond the original psychological models, the authors next extracted a new set of dimensions by combining the unique variance captured by each of the prior models, using their Internet ratings together with dimensionality reduction (principal components analysis; Fig. 1*B*). They found that four orthogonal dimensions captured most of the variance in the ratings accounted for by the seven models. These dimensions were linear combinations of the original 16 dimensions, and the authors gave them names based on their respective loadings: "rationality," "social impact," "human mind," and "valence." They tested this new set of dimensions against the neural data, one-by-one, and when they looked at neural similarity derived from the social brain network, rationality, social impact, and valence explained significant variance, whereas human mind did not. Tamir et al. (1) conclude that these three dimensions organize our thinking about other people's minds, and they estimate that this 3D space explains nearly 50% of the variance in the neural responses (averaged across subjects) to the 60 mental states measured in the fMRI study.

This new study is exceptionally thorough, but also exceptionally complex, and the final conclusions leave us with more questions than answers. It is somewhat perplexing that, even with the best combination of existing psychological models (the four principal components model introduced by the authors), the results only explain slightly less than 50% of the variance in mental representations [explained variance is quantified again with another variant of RSA, slightly closer to the classical implementation (6), in *Supporting Information* of the study by Tamir et al. (1)]. Clearly, there is a whole lot more to be understood about how the brain represents the landscape of mental states. One likely reason for the low total variance accounted for is that the study completely discards individual differences: It is likely that the landscape I have built through my experiences to make sense of the minds of others is different from the one you have built. Although these individual differences may be reflected in the neural data, they are not yet exploited in the models, which are based on average ratings from an independent sample of subjects. A next step that could, in principle, be taken on the extant data would be to parse the individual variability in neural similarity matrices, and thus extract new, perhaps subject-specific, dimensions from the neural data.

Relatedly, the authors speculate that the dimensions they found reflect aspects of social behavior that need to be solved as we interact with others, which seems plausible. Indeed, the core problem that the brain is trying to solve here is how to predict the behavior of others through these representations (7). However, one wonders if the solution to this problem might be quite

culturally specific. Would people from very different cultures, who have different ways of thinking about mental states, show different representations in their brains? It would also be interesting to extend the investigation to individuals with psychiatric illnesses like autism who appear to have atypical social cognition: What are their neural representations of mental states?

Another intriguing further question concerns how the brain's representations of other minds might be related to representations of one's own mind. Although the evidence we have available for inferring mental states seems to differ radically in the two cases (we need to observe other people, but not ourselves, to know what they and we think, feel, and believe), it does seem as though all of the 60 mental state terms used in the study would apply equally well to thinking about one's own mind. Does their representation look the same in the brain?

A fuller description of how the brain represents mental states would need to trace the flow of information from social perception [seeing people behave in certain ways, known to be represented in specific regions of association cortex (8)] to what is presumably a series of neural representations that comprise the transformation from perception to the inference of mental states. The representations revealed in the present study are explicit representations that correspond to how we talk and think about mental states. Out of what building blocks are these dimensions constructed in the brain (9)? How do they, in turn, feed into valuation and action representations that ultimately determine our social behavior toward others? The authors, perhaps wisely, stop short of these considerations in their already ambitious paper, but the next generation of experiments should begin to embed the present findings into a more complex circuit from social perception to social action.

1 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.
2 Kriegeskorte N, Mur M, Bandettini P (2008) Representational similarity analysis - connecting the branches of systems neuroscience. *Front Syst Neurosci* 2:4.
3 Skerry AE, Saxe R (2015) Neural representations of emotion are organized around abstract event features. *Curr Biol* 25(15):1945–1954.
4 Stanley DA, Adolphs R (2013) Toward a neural basis for social behavior. *Neuron* 80(3):816–826.
5 Schurz M, Radua J, Aichhorn M, Richlan F, Perner J (2014) Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neurosci Biobehav Rev* 42:9–34.
6 Nili H, et al. (2014) A toolbox for representational similarity analysis. *PLOS Comput Biol* 10(4):e1003553.
7 Koster-Hale J, Saxe R (2013) Theory of mind: A neural prediction problem. *Neuron* 79(5):836–848.
8 Deen B, Koldewyn K, Kanwisher N, Saxe R (2015) Functional organization of social perception and cognition in the superior temporal sulcus. *Cereb Cortex* 25(11): 4596–4609.
9 Schaafsma SM, Pfaff DW, Spunt RP, Adolphs R (2015) Deconstructing and reconstructing theory of mind. *Trends Cogn Sci* 19(2):65–72.