Published in final edited form as:

Trends Cogn Sci. 2018 March ; 22(3): 201–212. doi:10.1016/j.tics.2017.12.005.

Modeling the predictive social mind

Diana I. Tamir* and Mark A. Thornton*

Department of Psychology, Princeton University, Princeton, NJ 08540, USA

Abstract

The social mind is tailored to the problem of predicting other people. However, social cognition researchers have only just scratched the surface of the predictive social mind. Here we discuss a new framework for explaining how people organize social knowledge and use it for social prediction. Specifically, we propose a multi-layered framework of social cognition, where two hidden layers – others' mental states and traits – support predictions about the observable layer – others' actions. A parsimonious set of psychological dimensions structures each layer, and proximity within and across layers guides social prediction. This simple framework formalizes longstanding intuitions from social cognition, and in doing so, offers a generative model for deriving new hypotheses about predictive social cognition.

From Social Knowledge to Social Prediction

To successfully interact with other people, we must anticipate their thoughts, feelings, and actions. Imagine the difficulty of navigating the social world if you could not anticipate that tired people tend to become frustrated, or that agreeable people tend to cooperate. Our social interactions depend on our capacity for social prediction, and our social predictions are predicated on knowledge about other people, such as their mental states (e.g., tired) or traits (e.g., agreeable). Lacking such insight might result in a social life filled with faux pas, miscommunication, and missed opportunities. Fortunately, the social mind appears tailored to the problem of predicting other people. Our occasional lapses only highlight the accuracy and automaticity of everyday social cognition. As of yet, however, cognitive science has only scratched the surface of the predictive social mind [1].

In the present work, we offer a perspective that addresses two key challenges facing social cognition research. First, we discuss what information people use to make social predictions. Is there a computationally efficient yet functional solution that allows people to represent the richness and complexity of other minds? Second, we discuss how people leverage these representations to predict others' behavior. Much previous research has focused on static emotions, but much less is known about emotional dynamics [2–4], or how those states

Correspondence: dtamir@princeton.edu & mthornto@princeton.edu.

*Equal author contributions

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

evolve into social behavior. How do people model the probabilistic connections from static states or enduring traits to make predictions about future states or actions?

To address these challenges, we propose a multi-layered framework for social cognition (Figure 1). This framework characterizes both the structure and dynamics of social representations, and in doing so, offers coherent, generalizable responses to the two challenges identified above. This framework comprises (at least) three layers: the first layer describes others' observable actions; and two additional hidden layers concern others' mental states and traits, respectively. The mental state layer describes the thoughts, feelings, and perceptions – such as joy, exhaustion, planning, and contemplation – that define an individual's internal mental life. The trait layer describes reliable individual differences in social identity: stable personality attributes, such as being agreeable, intelligent, or optimistic. Each layer in this model is defined by a low-dimensional structure [5, 6]; this structure reflects how people distill each nuanced mental state or person down to coordinates on a parsimonious set of psychological dimensions. A map of this dimensional structure thus offers an efficient coding scheme for how people organize information about others' states and identities.

We propose that people map not just the locations of each trait, state, and action in each layer, but also probabilities of transition from one point to another [7]. These probabilistic trajectories within, and between, layers offer a simple explanation for how people might use their social knowledge to predict others' futures. If the location of a mental state or trait implies its trajectory, one could predict states from traits, and actions from states. This would largely solve the social prediction problem: people could predict where in social space another person will go, based on where that person currently resides.

Here we summarize evidence for these propositions, derived using modern methods from neuroimaging and machine learning. We also offer suggestions for how this framework might be extended, and how it may reframe longstanding issues in theory of mind.

Social Knowledge: Mapping the Structure of Each Layer

The first challenge facing predictive social cognition is to understand what information people might draw on to make social predictions. How do the mind and brain represent the richness and complexity of other people? We propose that the human mind simplifies this problem by organizing social information in relatively low-dimensional representational spaces. That is, people do not represent every nuance of others' mental states and traits; instead people deal with the complexity of others' minds by focusing on the positions of those states and traits on just a few key dimensions. This dimensionality reduction alleviates the burden of having to consider every minute detail of another person and their internal experiences, while still capturing the essential information in a precise map. Here we discuss research that uses recent advances in neuroimaging techniques to map the dimensional structure of these hidden layers.

The structure of the trait layer

Decades of methodical work in the field of personality psychology have identified multiple dimensions that capture people's personalities. Personality taxonomies that try to capture actual individual differences in enduring behavioral tendencies, such as the "Big Five" model [8, 9], have identified the following dimensions: extraversion, narcissism, conscientiousness, agreeableness, and openness. Similarly, theories from person perception have identified six additional dimensions: warmth and competence [10, 11], trustworthiness and social dominance [12], and agency and experience [13]. Each of these dimensional theories successfully describes the phenomena to which it is tailored – personality, stereotypes, face perception, mind perception, respectively. However, any individual has many more features than can be captured by just these few dimensions. When making social predictions, do people indeed rely on just a few dimensions, such as those identified in these theories of personality and person perception?

In recent research [14], we used representational similarity analysis (RSA; see Box 1), and voxelwise encoding models (see Box 2) to simultaneously test, compare, and synthesize the existing theories of trait space. These analyses focused on activity within the neural network that encodes social knowledge. This network includes dorsomedial prefrontal cortex (dMPFC), ventromedial prefrontal cortex (vMPFC), medial parietal cortex (MPC), temporoparietal junction (TPJ), and the anterior temporal lobes (ATL). Univariate analyses have demonstrated that this network reliably responds to many varieties of social content, including thinking about mental states, making inferences about others' beliefs, considering individual people and personalities, or groups of people and stereotypes [15–17]. By analyzing distributed patterns of brain activity within this neural network, we could determine which trait dimensions people spontaneously encode during social inferences. Results showed that people employ three dimensions to map others' traits: power, valence, and sociality. These dimensions combine to make a synthetic model that integrates insights from each of the existing dimensional theories: power loads heavily on the dimensions dominance, conscientiousness, and agency; valence loads on warmth, trustworthiness, and agreeableness; and sociality loads primarily on extraversion. Together, these three dimensions explain approximately two-thirds of the reliable variance in neural activity elicited by thinking about different people, thereby offering the most comprehensive model of trait representation to date. Thus, much of the richness of others' minds can indeed be compressed to coordinates in a low-dimensional trait space. The approach employed here concurrently enriches our knowledge about what content the social brain encodes, while also helping to refine existing psychological models.

The structure of the mental state layer

Just as people must make sense of others' enduring traits, people must also make sense of others' momentary mental states. Again, we propose that the mind and brain do so by arraying mental states within a low dimensional representational space. Decades of research in social psychology offer at least seven well-supported theories about the psychological dimensions that people encode when considering others' mental states. These dimensions include valence and arousal [18, 19], warmth and competence [10, 20], agency and experience [13], emotion and reason, mind and body [21], social and nonsocial [15, 22, 23],

and uniquely human and shared with animals [24]. Can just a few such dimensions explain how people think about others' mental states?

Once again, we applied RSA to functional neuroimaging data to test and synthesize these theories [5]. Our results indicate that people are indeed attuned to a small set of orthogonal dimensions when thinking about others' mental states: rationality, social impact, and valence (Figure 2). These dimensions were derived using principal components analyses over the dimensions of the seven existing social psychological theories listed above: rationality loaded highly in one direction on the dimensions emotion, experience, and warmth, and loaded highly in the opposite direction on reason, agency, and competence; social impact loaded positively on the social and high-arousal dimensions, and negatively on nonsocial and low-arousal; valence loaded positively on positive and warmth, and negatively on negative. This three-dimensional map accounts for almost half of the reliable variation underlying neural patterns within the neural network that support social cognition, and thus offers the most predictive model of mental state representations to date [5]. To put these results in context, the neural variance explained here is comparable to that explained by models in more computationally tractable domains such as vision [25]. We predict that these same dimensions the brain uses to encode mental states will also facilitate social prediction. If so, this would speak to the foundational role of these particular dimensions in solving both major challenges in social cognition.

The structure of the action layer

The action domain is, in many ways, the ultimate test of social prediction. Predictions about others' thoughts and feelings, as well as reverse-inferences about their traits, are valuable in the social world. However, to maximally benefit from knowledge of traits and states, people must ultimately "cash out" this knowledge in the hard currency of behavioral predictions. Additionally, since actions are more perceptible, both first-and second-hand, than mental states or traits, taking in information at the action level is likely crucial to forming an accurate mental model of the social world.

Unfortunately, the action layer remains largely terra incognita. There are many more discrete actions than there are states or traits, and moreover, each individual action can be described at many different levels. For instance, one could label the same physical action as kicking, passing, playing soccer, exercising, or cooperating. While systems neuroscientists might study the execution of simple actions like reaching in non-human primates [26], psycholinguists might study how children use the syntactic properties of words that represent actions (i.e., verbs) to learn their meaning [27]. A socially-relevant taxonomy of action would likely fall somewhere between these extremes. We suggest that a successful characterization of the action layer will capture socially-meaningful, real-world features of these action, such as who is likely to perform the action (e.g., by powerful, positive, social people, or powerless, negative, asocial people), when and where the action is likely to be performed (e.g., morning vs. night or inside vs outside), and how the action is performed (e.g., mentally or physically). To date, one of the most studied dimensions of why people act has been that of "fight versus flight," often referred to as approach versus avoidance [28].

We speculate that the mind might use this, and other dimensions, such as valence or costliness, to evaluate others' actions.

One approach for uncovering the dimensions people employ for thinking about actions will be structural analysis of verbs – linguistic proxies for actions – and their use in natural language. Existing word taxonomies and models of language, such as wordnet and word2vec, offer a benchmark for any future psychologically refined model of actions. Moreover, with the advent of the new analytic techniques described above, computational fMRI may also be ideally placed to contribute to this theory-building process. We look forward to research that identifies, evaluates, and winnows the dimensions organizing actions into an efficient representational space.

Social Prediction: Mapping Transitions Within and Between Layers

The second major challenge for social cognition is understanding how people leverage their social knowledge to make social predictions. The structure of state representation discussed above has thus far been defined by how people think about another person at a single moment in time. However, people are dynamic [4]. States unfold over time, each with their own variable temporal profile; for instance, surprise ends quickly, but sadness can extend indefinitely. States also transition seamlessly from one into the next; you may be hungry one moment, but that hunger may soon beget frustration and anger, lest you eat some food. Navigating social life depends on understanding these dynamics and using them to predict other people.

The brain is well suited to make such predictions [29–32]. The theory of predictive coding suggests that the brain represents the world not as percepts, but as predictions [33–36]. In the domain of perception, people are known to make these automatic predictions [34, 37]. For example, when one sees a ball in flight, one reflexively makes a prediction about its trajectory. People use contextual cues from the environment to predict where objects are likely to appear (e.g., a baseball glove at a baseball stadium) or which events are likely to occur (e.g., the national anthem at a baseball game). People also make reflexive predictions in cognitive domains [38, 39]. For example, when processing language, one might use the beginning of a sentence to predict the end of that seamstress [39]. We also use prediction errors – differences between the prediction and the observation (e.g., “sentence” vs. “seamstress”) – to shape subsequent predictions (See Box 3).

The same may be true in the domain of social cognition: people might use their knowledge of others' mental states and identities to predict others' actions. When seeing a friend fly into a rage, one might reflexively predict that person's emotional trajectory, and the action upon which they might land. We propose that people make such predictions by modeling the transitional probabilities between points both within and across the layers of social cognition. That is, people leverage knowledge of the psychological locations and relations between traits, states, and actions to predict the social future. In doing so, people intuit how traits predict mental states (e.g., agreeable people tend to feel content), how mental states predict other mental states (e.g., content people tend to later feel grateful) [7], how mental states predict actions (e.g., grateful people tend to cooperate) [26–28], and how traits predict

actions (e.g., agreeable people tend to cooperate) [29]. In each of these cases, people use knowledge about one layer to make predictions, eventually moving toward the end goal of predicting observable actions. Social cognition research has long assumed that people make these intuitive predictions between traits, states, and actions. However, little work has taken on the challenge of formalizing how people implement these predictions.

Mapping transitions within the mental state layer

Two prerequisites must be met for people to be able to make useful predictions within the mental state layer. First, mental states must transition from one to another with regularity. That is, a person's current mental state must actually predict their future mental state(s). Second, perceivers must have accurate mental models of these regularities in state transitions. For example, if one can see that a colleague is currently tired, and one has the intuition that tiredness leads to frustration, then one could make a useful prediction that the colleague may later feel frustrated – but only if tiredness actually precedes frustration with some regularity.

Recent research suggests that mental state predictions meet both prerequisites [7]. First, experience-sampling studies indicate that mental states transition from one to the next in reliable ways. Second, people have accurate judgements of the likelihoods of these transitions; people's mental models do indeed reflect statistical regularities in others' affective dynamics. Indeed, these mental models are sufficiently accurate to predict not just the next state, but two highly specific states into the future. Thus, a combination of accurate direct social perception in the moment, with accurate mental models of state transitions in the perceiver's mind, may indeed allow perceivers to predict others' future states. Although such predictions might never be certain, they could help social prognosticators constrain the probability distribution over likely social futures, and act accordingly.

The structure of mental state representation shapes state transitions. The closer two states are on the dimensions of rationality, social impact, and valence, the more people predict that others will likely transition between those states [7]. For instance, gratitude and joy are both high impact, positive, emotional states, and therefore close together in state space. As a result, people are more likely to predict transitions from gratitude to joy than to a psychologically distant emotion like contempt. Proximity on these three dimensions also predicts actual transitional probabilities. Indeed, the structure of the dimensional space mediates the accuracy of transition judgments. Thus, the psychological dimensions identified above serve as a scaffolding for social prediction.

People's judgements of the likelihood of state transitions may be highly accurate, but they are also somewhat egocentric [7]. People's own idiosyncratic emotional experiences impact their judgments about others' likely emotional transitions. Indeed, there is substantial variance in individuals' level of accuracy, even within the typical adult population thus far tested. The accuracy of people's models of emotion transitions might provide a useful assay of real-world social ability by capturing meaningful variance in social ability. As such, we suggest that researchers may be able to use this measure to test new hypotheses about typical adults, clinical disorders, and developmental populations. For example, given that one's own emotional experiences inform one's social predictions, individuals who experience aberrant

states or transitions might rely on this misleading source of information, and thus exhibit inaccurate predictions about others' state transitions. Thus, individuals with mood disorders might suffer from a double jeopardy: the mood disorder itself, and the social inaccuracy caused by their atypical affect. More generally, psychiatric disorders might be effectively characterized by disordered state transition experiences, rather than just an over- or under-abundance of particular states [40, 41].

Mapping transitions between layers

A full model of social prediction must map not only transitions within a layer, but also transitions between layers. How does information at one layer facilitate prediction at another? If you know the traits someone exhibits (e.g., high power & high negativity), then you should be able to predict their likely coordinates in state space (e.g., high social impact & negative valence), which you should be able to use to predict the type of action they're likely to take (e.g., an aggression). That is, probabilistic predictions should be used to constrain predictions from the trait to state layer, and from the state to action layer.

Researchers have already begun mapping the transitions from traits to states. The dimensions that structure the trait layer – power, valence, sociality – both conceptually and empirically mirror those that structure the state layer – rationality, valence, social impact. Models trained to predict personality traits can likewise predict others' momentary mental states [14]. This suggests that there is a partial overlap between the trait and states layers, such that people with certain traits (e.g. trustworthiness) systematically resemble people with particular states (e.g. happiness) in our minds. Indeed, it is possible that traits and states are not independent concepts, but instead sample much of the same information, simply in different temporal windows. Future neural evidence for such cross-layer transitions might come from pattern analysis – to determine whether pattern similarity reflects transitional probabilities – or from fMRI repetition suppression, where anticipated states should elicit less activity than unanticipated ones in regions that spontaneously encode the mappings within or between layers of social cognition.

Future work should also endeavor to map the transitions from states to actions. Predicting actions is the end goal of the predictive coding account of social cognition proposed here. This goal should become more feasible as researchers uncover the structure of the action layer.

Concluding Remarks

Here we have proposed a multilayered framework for social cognition. This framework is designed to explain how people represent other people's minds, and how people predict others' states and behaviors. First, we demonstrate that a parsimonious set of dimensions scaffolds both psychological and neural representations of others' traits and states [6, 14, 42]. This simple structure offers a computationally efficient coding scheme for how people represent social knowledge, and in doing so helps to unify multiple existing dimensional theories from the psychological literature. Second, we demonstrate how this social knowledge is used in service of making social predictions [7]. By placing prediction as the central goal of social cognition, we offer an additional *raison d'être* for many data-driven

theories of social content: the dimensions of such theories facilitate social prediction, by scaffolding an interconnected framework for the psychological organization of social content.

The multilayered framework formalizes many longstanding intuitions from the social cognition literature. It provides a natural way to translate social knowledge and social prediction – inherently “fuzzy” topics [22] – into precise mathematical language, through the use of metric representational spaces and transitional probabilities within and between them. Such formalisms in turn allow for the generation and testing of specific numerical hypotheses, upon which new and improved theories may be built. We also suggest that this framework can also be used to help resolve longstanding debates within social cognition (Box 4), such as the potentially false dichotomy between simulation theory and theory theory.

While this framework aims to provide a productive foundation for future research on predictive social cognition, there remain multiple notable gaps in this framework’s explanation for how people implement social predictions in the real world (see Outstanding Questions Box). First, researchers have yet to map the structure of the action layer, or any transitional probabilities to, from, or within it. We offer some suggestions above for how to approach this task, and what an ideal solution might look like.

Second, we have yet to account for the considerable role that situation and context play in reshaping the structure or dynamics of the framework. The vast social psychological literature on situation-dependence of personality traits suggests that situations can predict behavior more powerfully than traits [43–45]. Our initial focus on the trait layer reflects people’s tendency to over-rely on dispositional tendencies to predict and explain behavior while neglecting the influence of social situations on behavior – thus succumbing to the fundamental attribution error [46, 47]. Nevertheless, situations exert enormous influence on people’s behavior, and people do indeed take context into account when making social predictions [48]. As such, a complete model of social prediction must include an account for how people use situations to predict people’s states and behavior.

Third, the predictions of our model should be tested in dynamic, naturalistic environments. The simple, static stimuli and generic targets employed to develop the model represent a restricted range of the stimuli people encounter in the social world. Although it is noteworthy that the current framework can explain robust neural variance and behavioral transition predictions even in the absence of cues about the person, event, and situation, future research should test and expand this model using a wider range of ecologically valid stimuli, and individuated, familiar targets.

We see the current framework as integrating well-established social psychological theories with newer neural and computational methods to understand real-world social functioning. Success in the social world requires people to make predictions about others’ actions by combining information about other’s personality, mental state, and situation from multiple noisy sources. Here we sketch a framework which describes how people might represent this information, and use it in service of social predictions. We hope to encourage future

researchers both to employ this framework productively to model real-world social predictions, and constructively, to enhance its predictive potential.

Acknowledgments

The authors thank Ida Momennejad, Lily Tsoi, and Zidong Zhao for their feedback on earlier version of this manuscript, and Kathleen Cantner for her illustration of Figure 1. This work was supported by NIMH grant R01MH114904 to D.I.T.

References

1. Koster-Hale J, Saxe R. Theory of mind: a neural prediction problem. *Neuron*. 2013; 79(5):836–848. [PubMed: 24012000]
2. Heller AS, Casey B. The neurodynamics of emotion: delineating typical and atypical emotional processes during adolescence. *Developmental science*. 2016; 19(1):3–18. [PubMed: 26548573]
3. Schirmer A, et al. The socio-temporal brain: Connecting people in time. *Trends in cognitive sciences*. 2016; 20(10):760–772. [PubMed: 27615804]
4. Cunningham WA, et al. Emotional states from affective dynamics. *Emotion Review*. 2013; 5(4): 344–355.
5. Tamir DI, et al. Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences of the United States of America*. 2016; 113(1):194–199. [PubMed: 26621704]
6. Thornton MA, Mitchell JP. Consistent Neural Activity Patterns Represent Personally Familiar People. *Journal of Cognitive Neuroscience*. 2017; 29(9):1583–1594. [PubMed: 28557690]
7. Thornton MA, Tamir DI. Mental models accurately predict emotion transitions. *Proceedings of the National Academy of Sciences*. 2017; 114(23):5982–5987.
8. McCrae RR, Costa J, Paul T. Validation of the five-factor model of personality across instruments and observers. *Journal of Personality and Social Psychology*. 1987; 52(1):81–90. [PubMed: 3820081]
9. Goldberg LR. An alternative “description of personality”: the big-five factor structure. *Journal of Personality and Social Psychology*. 1990; 59(6):1216–1229. [PubMed: 2283588]
10. Cuddy AJ, et al. Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology*. 2008; 40:61–149.
11. Fiske S, et al. A model of (often mixed) stereotype content: competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*. 2002; 82(6):878–902. [PubMed: 12051578]
12. Oosterhof NN, Todorov A. The functional basis of face evaluation. *Proceedings of the National Academy of Sciences of the United States of America*. 2008; 105(32):11087–11092. [PubMed: 18685089]
13. Gray HM, et al. Dimensions of Mind Perception. *Science*. 2007; 315(5812):619. [PubMed: 17272713]
14. Thornton MA, Mitchell JP. Theories of Person Perception Predict Patterns of Neural Activity During Mentalizing. *Cerebral Cortex*. (In press).
15. Mitchell JP. Contributions of functional neuroimaging to the study of social cognition. *Current Directions in Psychological Science*. 2008; 17(2):142–146.
16. Saxe R, Kanwisher N. People thinking about thinking people: The role of the temporo-parietal junction in “theory of mind”. *NeuroImage*. 2003; 19(4):1835–1842. [PubMed: 12948738]
17. Contreras JM, et al. Dissociable neural correlates of stereotypes and other forms of semantic knowledge. *Social Cognitive and Affective Neuroscience*. 2012; 7(7):764. [PubMed: 21908447]
18. Posner J, et al. The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology. *Developmental Psychopathology*. 2005; 17(3):715–734.

19. Russell JA. A circumplex model of affect. *Journal of Personality and Social Psychology*. 1980; 39(6):1161–1178.
20. Fiske ST, et al. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*. 2002; 82(6):878–902. [PubMed: 12051578]
21. Forstmann M, Burgmer P. Adults are intuitive mind-body dualists. *Journal of Experimental Psychology: General*. 2015; 144(1):222–235. [PubMed: 25494547]
22. Mitchell JP. Social psychology as a natural kind. *Trends in Cognitive Sciences*. 2009; 13(6):246–251. [PubMed: 19427258]
23. Britton JC, et al. Neural correlates of social and nonsocial emotions: An fMRI study. *Neuroimage*. 2006; 31(1):397–409. [PubMed: 16414281]
24. Haslam N. Dehumanization: An integrative review. *Personality and Social Psychology Review*. 2006; 10(3):252–264. [PubMed: 16859440]
25. Kriegeskorte N, et al. Representational similarity analysis: Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*. 2008; 2(4):1–28. [PubMed: 18958245]
26. Wessberg J, et al. Real-time prediction of hand trajectory by ensembles of cortical neurons in primates. *Nature*. 2000; 408(6810):361–365. [PubMed: 11099043]
27. Naigles L. Children use syntax to learn verb meanings. *Journal of child language*. 1990; 17(2): 357–374. [PubMed: 2380274]
28. Elliot AJ, Thrash TM. Approach–Avoidance Motivation in Personality: Approach and Avoidance Temperaments and Goals. *Journal of Personality and Social Psychology*. 2002; 82(5):804–818. [PubMed: 12003479]
29. von Helmholtz, H. *Treatise on physiological optics*. Courier Corporation; 1910.
30. Gregory RL. Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*. 1980; 290(1038):181–197. [PubMed: 6106237]
31. Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*. 2013; 36(03):181–204. [PubMed: 23663408]
32. Friston K, Kiebel S. Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society B: Biological Sciences*. 2009; 364(1521):1211–1221.
33. Kilner JM, et al. Predictive coding: an account of the mirror neuron system. *Cognitive Processing*. 2007; 8(3):159–166. [PubMed: 17429704]
34. Rao RP, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*. 1999; 2(1):79–87. [PubMed: 10195184]
35. Saygin AP, et al. The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social cognitive and affective neuroscience*. 2012; 7(4):413–422. [PubMed: 21515639]
36. Van de Cruys S, et al. Precise minds in uncertain worlds: predictive coding in autism. *Psychological review*. 2014; 121(4):649–675. [PubMed: 25347312]
37. Vuust P, et al. Predictive coding of music–Brain responses to rhythmic incongruity. *Cortex*. 2009; 1(45):80–92.
38. Lupyan G, Clark A. Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychological Science*. 2015; 24(4):279–284.
39. Lewis AG, Bastiaansen M. A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. *Cortex*. 2015; 68:155–168. [PubMed: 25840879]
40. Eldar E, et al. Mood as representation of momentum. *Trends in Cognitive Sciences*. 2016; 20(1): 15–24. [PubMed: 26545853]
41. Eldar E, Niv Y. Interaction between emotional state and learning underlies mood instability. *Nature Communications*. 2015; 6
42. Tamir DI, et al. Neural evidence that three dimensions organize mental state representation: Rationality, social impact, and valence. *Proceedings of the National Academy of Sciences*. 2016; 113(1):194–199.

43. Asch SE. Studies of independence and conformity: I. A minority of one against a unanimous majority. *Psychological monographs: General and applied*. 1956; 70(9):1.
44. Milgram S. Behavioral Study of Obedience. *Journal of Abnormal and Social Psychology*. 1963; 67:371–378.
45. Darley JM, Batson CD. from Jerusalem to Jericho. *Journal of Personality and Social Psychology*. 1973; 27(1):100–108.
46. Gilbert DT, Malone PS. The Correspondence Bias. *Psychological Bulletin*. 1995; 117(1):21–38. [PubMed: 7870861]
47. Jones EE, Harris VA. The attribution of attitudes. *Journal of Experimental Social Psychology*. 1967; 3(1):1–24.
48. Ross L, Nisbett RE. *The person and the situation: Perspectives of social psychology*, Pinter & Martin Publishers. 2011
49. Todorov A, et al. Toward socially inspired social neuroscience. *Brain research*. 2006; 1079(1):76–85. [PubMed: 16480692]
50. Kriegeskorte N, Kievit RA. Representational geometry: integrating cognition, computation, and the brain. *Trends in cognitive sciences*. 2013; 17(8):401–412. [PubMed: 23876494]
51. Kriegeskorte N, et al. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*. 2008; 2
52. Skerry AE, Saxe R. Neural Representations of Emotion Are Organized around Abstract Event Features. *Current Biology*. 2015; 25(15):1945–1954. [PubMed: 26212878]
53. Haxby JV. Multivariate pattern analysis of fMRI: the early beginnings. *Neuroimage*. 2012; 62(2): 852–855. [PubMed: 22425670]
54. Mitchell TM, et al. Predicting human brain activity associated with the meanings of nouns. *Science*. 2008; 320(5880):1191–1195. [PubMed: 18511683]
55. Huth AG, et al. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*. 2012; 76(6):1210–1224. [PubMed: 23259955]
56. Baker CL, et al. Action understanding as inverse planning. *Cognition*. 2009; 113(3):329–349. [PubMed: 19729154]
57. Saxe R, Houlihan SD. Formalizing emotion concepts within a Bayesian model of theory of mind\$. *Current Opinion in Psychology*. 2017; 17:15–21. [PubMed: 28950962]
58. Friston KJ, Frith CD. Active inference, communication and hermeneutics. *Cortex*. 2015; 68:129–143. [PubMed: 25957007]
59. Gordon RM. The simulation theory: Objections and misconceptions. *Mind & Language*. 1992; 7(1-2):11–34.
60. Carruthers, P., Smith, PK. *Theories of theories of mind*. Cambridge University Press; 1996.
61. Schacter DL, et al. Episodic simulation of future events. *Annals of the New York Academy of Sciences*. 2008; 1124(1):39–60. [PubMed: 18400923]

Box 1

Testing Psychological Theories with Representational Similarity Analysis

In the early years of social neuroscience, there was a disconnect between the questions which psychology had historically posed (e.g. “which trait dimensions define personality?”), and the types of questions human neuroimaging was trying to answer (e.g., “which brain regions respond more to social than non-social stimuli”) [49]. As a result, much of the early psychological progress made by fMRI relied on associating or dissociating process in the brain [15], while theories of mental content organization often went untested. Now, however, nascent neuroimaging analysis methods permit direct tests of existing social psychological theories. Representational similarity analysis (RSA) is one such technique which can be used to arbitrate amongst hypotheses generated by different psychological theories [50, 51]. RSA examines the patterns of neural activity elicited by different stimuli; it compares actual neural pattern similarity to the similarity predicted by a theory.

RSA has recently been used to test which dimensions people use to represent mental states [5, 52]. To do so, researchers first measured distributed patterns of neural activity elicited by thinking about different mental states. For example, on one trial participants might see the word “love” and decide which of two situations were more associated with that mental state (e.g., “hugging your mom” and “writing a love letter”); a reliable pattern corresponding to the neural representation of “love” could be estimated by averaging over many such trials with varied situations. Next, researchers assessed the similarities and dissimilarities (i.e., correlations) amongst such patterns. For instance, how similar is the pattern for “love” to the pattern for “envy”? Finally, researchers compared those neural similarities to the similarities predicted by competing hypotheses. For example, the hypothesis that people use the social impact dimension to understand states would predict that the states of “love” and “envy” should elicit very similar patterns, since they are both similarly socially impactful. In contrast, the hypothesis that people use the dimension valence would predict that “love” and “envy” should elicit very different patterns because one state is positive, whereas the other is negative. Each psychological dimension makes predictions about the similarity of each mental state compared with every other mental state. RSA simultaneously assesses the accuracy of all such predictions by correlating neural pattern similarity with the predictions of similarity or difference made by each dimension. In doing so, RSA concurrently reveals which psychological dimensions shape patterns of neural activity, and which neural regions employ those dimensions.

Box 2

Building Encoding Models of Social Cognition

The advent of multivoxel, or multivariate, pattern analysis (MVPA) techniques have brought about significant advances in the capabilities of functional neuroimaging [53]. MVPA is an umbrella term that covers several related techniques – including RSA (Box 1) and decoding/classification – which analyze the activity of multiple voxels together, rather than in the massively parallel univariate approach of the traditional general linear model. Voxelwise encoding modeling is one such new form of MVPA. Encoding models test psychological theories by evaluating how well their dimensions can explain the activity of each voxel in the brain, as well as in the emergent patterns distributed across these voxels [54, 55].

Encoding models were recently used to test which dimensions the brain uses to represent others' traits [14]. In this study, encoding models first learned a mapping between the dimensions of a psychological theory (e.g., stereotype content [10, 11]) and voxelwise activity in the brain. To train a stereotype content encoding model, participants made repeated judgments about target people (e.g. how much do they “enjoy spending time in nature?” or “dislike travelling by airplane?”). These targets varied in their warmth and competence, as rated by a separate set of participants. The model “learned” the pattern of brain activity associated with thinking about a canonical warm person by averaging across the patterns elicited by thinking about the targets – with strong weights on the warm individuals and weak weights on the cold ones [14]. This encoding model was then used to predict the patterns of neural activity for “test” targets not included during training. To do so, it mixed together its canonical warmth and competence patterns, weighted by the target's levels of warmth and competence. The predicted patterns were then compared with the actual patterns elicited by the test targets to determine the model's accuracy. Accurate predictions provide strong evidence that a theoretical dimension supports the representation of individual targets. This analysis can also test which such organizing theories performed better or worse than others (e.g., stereotype content vs. Big Five), and how well any of these theories' performance compared to an estimate of ideal prediction, at both voxelwise and pattern-wise levels. These capabilities make encoding models an attractive tool for understanding social knowledge through neuroimaging.

Box 3

Social Prediction in a Bayesian Brain

The predictive framework proposed in this review can be naturally and productively integrated with both Bayesian models of social cognition [56–58] as well as predictive coding theory – a biologically plausible algorithm for implementing Bayesian cognition [33–36].

In our framework, traits, states, and actions are defined by their coordinates in low-dimensional psychological spaces. Coordinates in these spaces describe not only the static features of these social stimuli, but also the likely transitions between stimuli: transitions are more likely between emotions that are close in the state-space, like joy and friendliness, than between emotions that are far apart, like joy and bitterness. In Bayesian terms, the coordinates that represents another person’s current state sit at the center of a probability distribution over prior beliefs about that person’s future states (Figure 1). A smooth function (inversely) relates distance between coordinates in the space to the probability density of the prior distribution. In line with Bayesian updating, receiving new information can sharpen or broaden this function based on confidence, or move the center of the distribution. Representing priors thus requires only a set of coordinates and a probability function, rather than a full transitional probability matrix. As such, this low-dimensional representational space allows for greater computational efficiency under a Bayesian approach.

The multilayered framework we propose also mirrors a predictive coding implementation of Bayesian cognition, in at least two ways [33–36]: First, in predictive coding, representations automatically incorporate predictions. For example, when a person sees a ball in flight, they can’t help but predict its trajectory. Similarly, in our framework, when one thinks about a social stimulus, such as another person’s mental state, one can’t help but encode that person’s affective trajectory – the most likely future states are nearby in state-space. Second, in predictive coding, predictions are compared to new sensory input, and resulting prediction errors are used to adjust future predictions. In our framework, prediction errors can be calculated by taking vector differences between the coordinates of previously inferred and newly inferred trait, states, or actions. Such prediction errors could be added to the existing coordinates of social stimuli to adjust their positions in the representational space in a way that would minimize future errors. Repeating this process over development might offer an effective way to establish social knowledge and form these representational spaces.

Box 4

Synthesizing Simulation Theory and Theory Theory

Theory-theory (TT) and simulation-theory (ST) propose two divergent mechanisms by which people understand one another [59, 60]. TT proposes that people make use of explicit folk theories to understand others' thoughts and actions; ST proposes that people use mental simulations, built upon self-knowledge, to understand others¹.

Debates between these two mechanisms often stall upon disagreements about the content that forms the basis for social understanding (theory vs. subjective experience) and its source (observing the world vs. the self). Our framework proposes that there are actually important similarities in the content that both theories draw on, and they diverge primarily in the processes by which they draw on this content. Specifically, we suggest that both types of content shape people's predictive models of social cognition: that is, people learn the probabilistic connection between social stimuli both through observation and personal experience. You could learn that hunger leads to anger through observing friends get hangry, and by experiencing it yourself; prediction errors about either the self or other could subsequently shape future predictions (Box 3). As such, we suggest refocusing the debate between ST and TT on process. In that light, our framework offers an opportunity to reconcile these theoretical differences.

ST and TT employ two distinct processes to make social predictions. We propose that both processes draw upon the probabilistic connections between traits, states, and actions encapsulated in our multilayer framework, but that they do so in distinct ways. A TT approach would make predictions by explicitly reasoning through the most probable links. TT thus offers a direct, serial route to deriving the most likely conclusions; however, its focus is narrow, ignoring low probability alternatives. In contrast, an ST approach would draw on the probabilistic links by sampling. Rather than taking only the path of greatest likelihood, it takes random walks through the framework's transitional probabilities – algorithmically mirroring a Markov Chain Monte Carlo simulation – with each step activating relevant episodic exemplars [61]. Across multiple walks, the statistically most likely paths will be frequently traversed, but lower probability paths will still be explored, forming a posterior distribution over outcomes. ST thus requires greater investment of computational resources, but its results are more comprehensive. We look forward to future research testing predictions derived from this proposed tradeoff.

¹Note that simulation theories can be cashed out in terms of both cognitive simulations and embodied simulation mediated by mirror neurons; here we focus on the former.

Trends Box

- We propose a multilayered framework consisting of two hidden layers – traits and states -and one observable layer – actions. This framework address two key challenges in social cognition: organizing social knowledge efficiently, and using it for social prediction.
- FMRI, combined with advanced analytic techniques such as representational similarity analysis and encoding models, offers a way to reveal the dimensional structure organizing each layer of social cognition.
- Three dimensions – rationality, social impact, and valence – organize the layer of mental states; three dimensions – power, sociality, and valence – organize the domain of traits. Cross-encoding analysis suggests that these layers may be partially overlapping.
- Proximity within state space predicts perceived and actual transitional probabilities between emotions, and mediates the accuracy of people’s perceptions. Thus, the organizational dimensions of social content may scaffold social prediction.

Outstanding Questions

- What is the structure of the action layer? What are the transitional probabilities between the trait layer and the action layer, between the state layer and the action layer, and within the action layer?
- Does the structure of the state layer remain stable across targets, or does it shift dynamically by target? For example, might the “human mind” dimension take on additional relevance when considering non-human animals or entities? Are people universally attuned to the rationality, social impact, and valence of a state, or do people flexibly tune into those dimensions depending on who is experiencing them?
- How can the current framework account for asymmetric transition probabilities? The current representational spaces use metric distance to predict transitions. However, people’s actual transitions are sometimes asymmetric, and people’s intuitions about others’ transitions are similarly asymmetric.
- How do people learn the structure of each layer across development? Do people develop the same dimensional structures across different cultures, or different languages? Do the representational structure and dynamics described here apply outside of WEIRD societies?
- How do situations, social groups, and dyadic relationships modulate the shape of the layers, or the relationships between them?

Figure 1.

The multilayer framework of social cognition. The model comprises one observable layer – actions – and two hidden layers – traits and states. Each layer is defined by a parsimonious set of psychological dimensions (axes). The state and trait layer have parallel structures: the dimensions Rationality, Social Impact and Valence organize the state layer, and Power, Sociality, and Valence organize the trait layer. Individual traits, states, and actions (clouds) are defined by their location within each layer. For example, “friendliness” is defined by its high emotionality, high social impact, and positive valence. Probabilistic predictions are made by mapping transitions (arrows) from one location to the next. Transitions are more likely between close points (dark arrows) than far points (light arrows); transitions can be asymmetric with respect to particular dimensions (e.g., one may be more likely to transition from high energy to low rather than vice versa). Transitions between layers are determined by how the dimensions from one layer map onto the dimensions of the next; for example, people predict that a person with positive traits is more likely to exhibit positive states than

negative states. Here we depict predictions from traits to states and traits to actions, from states to states and states to actions, and from actions to actions and to states.

Figure 2.

The structure of the mental state layer. The neural representation of mental states is defined by three psychological dimensions: social impact (y-axis), valence (x-axis), and rationality (orange gradient). Here we plot 60 mental states that span each of the three dimensions, employed in earlier research [42]. Positions on each dimension were derived from a principal components analysis of ratings of mental states on dimensions from existing psychological theories including valence and arousal [18, 19], warmth and competence [10, 20], agency and experience [13], emotion and reason, mind and body [21], social and nonsocial [15, 22, 23], and uniquely human and shared with animals [24]. Rationality loads highly in one direction on the dimensions emotion, experience, and warmth, and loaded highly in the opposite direction on reason, agency, and competence; social impact loads positively on the social and high-arousal dimensions, and negatively on nonsocial and low-arousal; valence loads positively on positive and warmth, and negatively on negative.