Available online at www.sciencedirect.com
ScienceDirect
ELSEVIER
Current Opinion in Psychology

Computational neuroscience approaches to social cognition

Leor M Hackel[1] and David M Amodio[2,3]

How do we form impressions of people and groups and use these representations to guide our actions? From its inception, social neuroscience has sought to illuminate such complex forms of social cognition, and recently these efforts have been invigorated by the use of computational modeling. Computational modeling provides a framework for delineating specific processes underlying social cognition and relating them to neural activity and behavior. We provide a primer on the computational modeling approach and describe how it has been used to elucidate psychological and neural mechanisms of impression formation, social learning, moral decision making, and intergroup bias.

Addresses
[1] Department of Psychology, Stanford University, Jordan Hall, 450 Serra Mall, Stanford, CA 94305, USA
[2] Department of Psychology, New York University, 6 Washington Place, New York, NY 10003, USA
[3] Department of Psychology, University of Amsterdam, Nieuwe Achtergracht 129, REC G, 1001 NK Amsterdam, NL

Corresponding authors: Hackel, Leor M (lhackel@stanford.edu), Amodio, David M (david.amodio@nyu.edu)

Current Opinion in Psychology 2018, **24**:92–97
This review comes from a themed issue on **Social neuroscience**
Edited by **David Amodio** and **Christian Keysers**

https://doi.org/10.1016/j.copsyc.2018.09.001

How do we form impressions of other people? Is Jane kind? Do we like her? And can we predict how she will act? These are core questions of social cognition—the field of psychology devoted to understanding the processes through which we perceive, represent, and act towards persons and groups. Social psychologists have pursued these questions for over 40 years, and the earliest social neuroscience studies probed the neural basis of impression formation and social attitudes. Recently, social neuroscientists have used computational approaches to advance and, in some cases, reconceptualize thinking on social cognition. Here, we provide a brief introduction to the computational modeling approach and highlight recent studies that have used it to elucidate social cognition.

Computational modeling approaches to social cognition and social neuroscience

The central aim of social cognition is to understand social behavior by elucidating its underlying cognitive and neural mechanisms. In the past, this was accomplished with careful experimentation and behavioral modeling (e.g. Process Dissociation, Quad Model) [1,2], but these approaches are limited in their ability to assess complex dynamic processes. Computational models allow researchers to probe trial-by-trial dynamics of learning and choice and to make precise quantitative predictions about behavior across time (Box 1). In neuroimaging research, computational models permit researchers to test neural correlates of theorized latent variables that are not directly observable in behavior. For example, in models of reinforcement learning, a key variable is *reward prediction error*—the discrepancy between the reward one receives and the reward one expected (see Box 1) [3]. By fitting behavior to reinforcement learning models, researchers can estimate a learner's trial-by-trial prediction errors. Fitting this timeseries to fMRI data can then identify neural regions that covary with prediction errors.

Formal models also permit researchers to compare human behavior to that of an optimal agent. For example, by comparing behavior and neural responses to a Bayesian model, one can ask whether people conform to rational principles of updating in social settings [4,5•] or deviate from rationality in systematic ways [6]. Such deviations may provide important clues to psychological and neural processes involved in behavior. Alternatively, with agent-based modeling, researchers can simulate agents that instantiate different models and identify which performs best in a given task (e.g. achieving the highest accuracy or winning the most money) [7••].

Neuroimaging provides clues about the cognitive processes that drive a behavior, and formal models offer a powerful approach to more precisely delineate such processes. For instance, people with stronger racial bias learn more readily to avoid threatening out-group faces [8]. This bias could emerge because they evaluate the faces more negatively (*differential evaluation*) or because they learn more efficiently (*differential learning*). Computational modeling supports the second explanation, providing insight into how social biases shape learning itself [8]. When combined with neuroimaging, formal models can identify dissociable patterns in neural activity, adding further precision in characterizing neurocognitive substrates of social behavior. Although one can never know whether one's theoretical account is the true explanation,

Box 1 A reinforcement learning modeling primer

Computational models provide an abstract mathematical description of how one might learn or make choices. Here, we offer a brief primer on *reinforcement learning* (RL) models, which have been influential in social neuroscience, although Bayesian and drift-diffusion models are also widely used.

RL models describe how an agent forms action preferences through trial and error. For instance, imagine choosing between two slot machines to earn money. Initially, you might choose randomly. As you win money, you form and update expectations about their respective payouts. In RL models, this update relies on a *reward prediction error*, or the difference between the reward one received and the reward one expected (symbolized δ):

$$\delta_t = Reward_t - Q_{t-1} \tag{1}$$

Here, Q_{t-1} represents the expected reward value as of the previous time point. Prediction errors are used to update one's previous estimate of reward upward or downward:

$$Q_t = Q_{t-1} + \alpha\delta_t \tag{2}$$

where α is a free parameter representing a learning rate. This parameter scales prediction errors and therefore controls the extent to which one updates expectations. With a learning rate of zero, agents would not update their expectations at all; with a learning rate of one, agents would fully update their expectations based on prediction errors.

Finally, an agent makes a choice on the subsequent trial, given the updated values of the slot machines. This is frequently modeled using a 'softmax' equation:

$$p_i^t = \frac{\exp(\beta \times Q_{i,t})}{\sum_j \exp(\beta \times Q_{j,t})} \tag{3}$$

where β is a parameter controlling stochasticity of choice, and $p_{i,t}$ is the probability of choosing option i (of j options) on trial t. This equation indicates that participants will probabilistically choose an option based on the difference in expected reward between them. Eq. (3) can be replaced with a linear function to model responses like reaction time [15] or skin conductance response [69].

Together, these equations specify how an agent learns which action to take through trial and error. Given a set of parameters α and β, these equations make quantitative predictions about how likely an agent would be to choose each option on each trial. Computer programs can find parameter values that maximize the match between predicted choices and actual choices. Moreover, alternate models can be specified and compared to see which provides the best match (for details, see Ref. [70]).

Finally, given best-fitting parameters, one can estimate the expected value or prediction error experienced by an agent on every trial during choice and feedback, respectively. This timeseries can serve as a regressor in fMRI analyses, identifying regions that show greater fMRI signal during trials with greater value or prediction error signals.

modeling can identify the best account among competing models (see also Refs. [9,10•]).

Contributions of computational approaches to social cognition & social neuroscience

Here, we review notable areas of innovation in computational social neuroscience, with special attention to the advantages described above: dissociating component processes, linking latent variables to the brain, gaining quantitative precision, and comparing behavior to optimality.

Social reinforcement learning

Humans tend to repeat actions that yield reward—a process known as reinforcement learning [3,11–13]. In non-social tasks (e.g. winning money from a slot machine), neural activity in ventral striatum correlates with reward prediction errors specified by computational models of reinforcement learning [14] (see Box 1). Computational studies reveal that similar processes also underlie social reinforcement learning, across behavior and the brain. First, social rewards—including smiling faces, social conformity, positive evaluations, and vicarious gains—can reinforce behavior in the absence of monetary reward [15–21]. Second, reward reinforcement processes can support the formation of social attitudes [22–24]. For example, when someone buys us lunch or pays a compliment, this rewarding feedback can encourage future interactions with that partner [7••]. Both kinds of social learning have been linked to reward prediction errors in ventral striatum, suggesting that similar computations underlie social and non-social reinforcement (but cf. Refs. [20,25,26].). As such, principles of reward learning can be brought to bear on sociocognitive questions, offering novel predictions about how people develop complex social attitudes and preferences [27•,28].

By comparison, traditional social psychological theories assume that attitudes are formed through passive observation of positive or negative information about a person and represented in a conceptual network [29,30]. The emerging evidence from reinforcement modeling suggests that these modes of attitude formation—conceptual and instrumental—may be complementary, and that each type of representation may support different aspects of attitude expression (e.g. in judgments and impressions as opposed to choice behaviors) [29]. Hence, a computational modeling approach promises to further illuminate the psychological mechanisms involved in the formation of social attitudes and preferences.

Impression formation

Computational models are shedding new light on how people update impressions over time. Prior fMRI investigations revealed that impression updates are associated with activity in a broad set of cortical regions, including dorsomedial prefrontal cortex (dmPFC), inferior parietal lobule (IPL), ventrolateral prefrontal cortex (vlPFC), and posterior cingulate cortex (PCC) [31–34]. These activations may reflect prediction errors related to the traits of others—that is, the discrepancy between a person's behavior and the behavior one expected based on a trait impression (e.g. of competence [6,35], generosity [7••,36], or trustworthiness [4,37–40]).

Computational studies suggest that distinct brain regions track two types of trait prediction errors. First, when observing others, people can update a conceptual representation of others' traits. Therefore, some studies have examined *absolute prediction errors*, or overall surprise, when people observe another's behavior—for instance, learning a target was more *or* less generous than expected [35,36]. These prediction errors correlate with cortical regions including those listed above. Second, when directly interacting with others, one can update value representations informed by another's traits, such as inferring that a generous person is a valuable interaction partner. Another line of work has therefore examined *directional prediction errors* during interaction, such as learning that a partner was more generous than expected [7••]. These prediction errors correlate with activity in ventral striatum—a region associated with reward and valuation—in addition to cortical regions noted above (Figure 1). Together, these studies reveal distinct ways in which the brain tracks the traits of others—one that may build a conceptual representation of others and one that may track the value offered by a partner's traits. In doing so, this work again expands the scope of social cognition to include impression formation through active social interaction and value-based learning.

Bayesian models may further reveal whether people update impressions in an optimal manner [4,6,7••,41]. For instance, Bayesian models specify how quickly a learner should revise prior beliefs in light of new evidence: in a *volatile* environment (featuring rapid change), one should update beliefs quickly in light of new evidence, whereas in a *stable* environment, one should update beliefs more slowly [42]. When learning about another person's trustworthiness, perceivers track the person's volatility and update impressions quickly or slowly as a result [4,41]. These estimates of volatility correlate with activation in the anterior cingulate cortex (ACC) gyrus [4]. Bayesian models can thus identify psychological and neural processes required for optimal impression updating.

By the same token, Bayesian models can reveal illuminating deviations from optimality. For example, when people learn about an advisor's competence, their judgments do not mimic an optimal Bayesian learner; instead, people show confirmation bias by down-weighting negative outcomes, thus maintaining unrealistic optimism about another's competence [6]. By providing a benchmark for rationality, Bayesian models can thus identify biases in impression updating.

Mentalizing

Social decisions often require a consideration of others' mental states. For instance, playing chess, giving gifts, and offering condolences require us to consider the intentions, preferences, and emotions of others. Computational studies characterize how people update their inferences about mental states and translate these inferences into choices. In the context of economic games, in which participants must often cooperate or compete with others, computational models specify how people update beliefs about the actions a competitor will take [43], try to influence others [44,45••] or reason about another's strategy [46].

Computational approaches can also address how we become more accurate in inferring others' mental states over time, such as learning that a friend often appears calm even when upset. This process can be modeled as a type of reinforcement learning: as people make judgments and receive feedback, they adjust their inferences by giving more or less weight to helpful or unhelpful cues (e.g. a calm facial expression) [47•]. Indeed, during feedback, reward prediction errors related to accuracy were found to correlate with ventral striatum activity. In contrast to traditional social psychological approaches employing single-shot judgments, this work suggests that people value social accuracy and improve their mental state inferences over time through reinforcement.

Observational learning

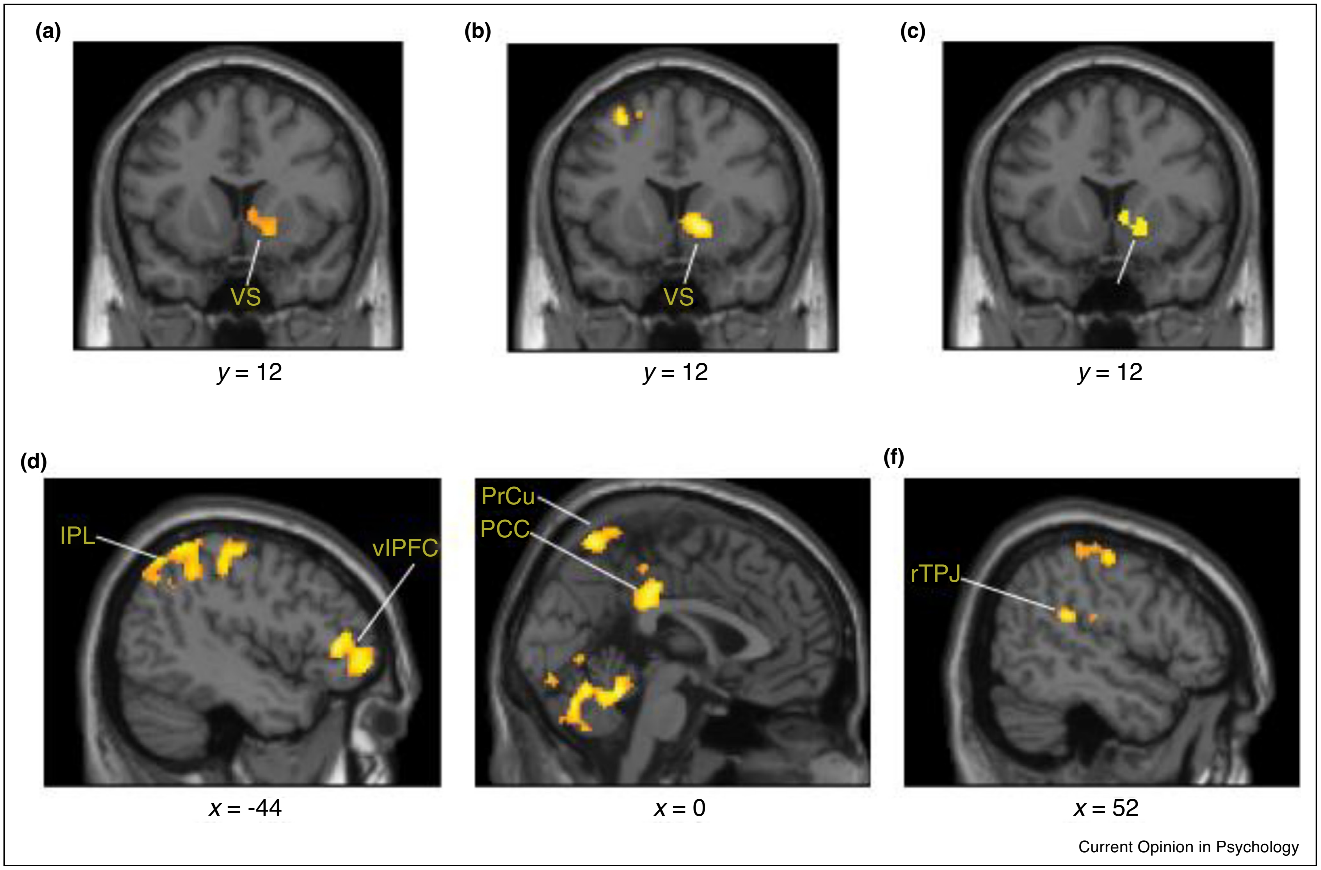
A key advantage of social living is that we learn from the mistakes and insights of others—a process known as observational learning [48]. Computational approaches can identify whether direct and observational learning rely on overlapping or distinct neural processes. During fear conditioning, a passive form of learning, similar neural computations support both processes [49]. In contrast, during instrumental learning, an active form of learning, different computations seem to support learning from experience and observation [50]. During both observational and direct instrumental learning, frontoparietal regions track whether an outcome is surprising; this response may reflect an abstract understanding of reward contingencies. During direct instrumental learning, ventral striatum further reflects whether an outcome is more rewarding than expected, suggesting the role of a second memory process [12,51,52].

Even without seeing others' outcomes, merely observing others' decisions can influence one's own choices [53–57]. Computational models allow researchers to test whether people use this social information optimally in light of their own uncertainty [5•] and to dissociate component processes underlying social influence [57]. Such studies thus reveal how social observation and direct experience are tracked and integrated in the brain.

Morality

People often must choose how to allocate gain or harm between oneself and others, and computational investigations connect these moral tradeoffs to the broader study

Figure 1



Neural correlates of prediction errors related to reward-based reinforcement and impression formation during social interaction, as revealed through computational modeling. Participants played an economic game in which partners varied in reward value (amount of money they shared) and generosity (proportion of available money they shared). Activity in ventral striatum (VS) correlated with **(a)** reward prediction errors and **(b)** generosity prediction errors; overlap shown in **(c)**. Notably, generosity prediction errors also correlated with activity in a set of regions previously associated with impression updating, including **(d)** ventrolateral prefrontal cortex (vlPFC) and inferior parietal lobule (IPL), **(e)** posterior cingulate cortex (PCC) and precuneus (PrCu), and **(f)** right temporoparietal junction (rTPJ). Reprinted from reference [7••].

of learning and choice [58]. Computational models can identify latent variables underlying these choices—such as the extent to which one values another's well-being or feels uncertain about another's preferences—and link these latent variables to brain activity [59,60,61••].

Computational studies have additionally brought new attention to the learning processes that give rise to moral judgments [62–64] and prosocial behavior [28,65,66]. For example, past work suggests that people reciprocate with individuals perceived to be generous [67,68]. Yet, reinforcement learning models suggest that people also like individuals who provide them with large material rewards [7••]. Indeed, people reciprocate more with wealthier partners who provide large material rewards, and this tendency correlates with the extent which people engage in reward-based learning [65]. This work thus reveals how learning dynamics give rise to morality.

Conclusion

The computational approach to social neuroscience offers important tools for elucidating the neural and cognitive processes that drive social cognition and behavior. This approach reflects a natural progression from documenting neural activations to probing their dynamic functions, and it is being applied to an expanding array of social processes—a harbinger of exciting innovations to come in the study of the social brain.

Conflict of interest statement

Nothing declared.

Acknowledgment

This work was funded by the US National Science Foundation (award 1606959). The authors acknowledge the generous support of the US National Science Foundation (1606959 to LH and 1551826 to DA) and the Netherlands Organisation for Scientific Research (VICI 016.185.058 to DA) during work on this article.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as

- • of special interest
- •• of outstanding interest

1. Amodio DM, Harmon-Jones E, Devine PG, Curtin JJ, Hartley SL, Covert AE: **Neural signals for the detection of unintentional race bias**. *Psychol Sci* 2004, **15**:88-93.

2. Conrey FR, Sherman JW, Gawronski B, Hugenberg K, Groom CJ: **Separating multiple processes in implicit social cognition: the quad model of implicit task performance**. *J Pers Soc Psychol* 2005, **89**:469.

3. Sutton RS, Barto AG: *Reinforcement Learning: An Introduction*. Cambridge: MIT Press; 1998.

4. Behrens TEJ, Hunt LT, Woolrich MW, Rushworth MFS: **Associative learning of social value**. *Nature* 2008, **456**:245-249 http://dx.doi.org/10.1038/nature07538.

5. • De Martino B, Bobadilla-Suarez S, Nouguchi T, Sharot T, Love BC: **Social information is integrated into value and confidence judgments according to its reliability**. *J Neurosci* 2017, **37**:6066-6074.

This study examined whether people follow tenets of Bayesian integration when relying on social consensus information (online product reviews). Participants primarily relied on product reviews when they felt uncertain and the reviews were reliable, in accordance with Bayesian principles, and dmPFC correlated with belief changes in light of these factors.

6. Leong YC, Zaki J: **Unrealistic optimism in advice taking: a computational account**. *J Exp Psychol Gen* 2018, **147**:170.

7. •• Hackel LM, Doll BB, Amodio DM: **Instrumental learning of traits versus rewards: dissociable neural correlates and effects on choice**. *Nat Neurosci* 2015, **18**:1233-1235 http://dx.doi.org/10.1038/nn.4080.

This study used computational modeling of reinforcement learning to examine trait-based impression formation, dissociating it from reward-based learning during social interaction. Modeling of prediction error signals revealed that while both forms of social learning depended on ventral striatum, trait learning further relied on additional cortical regions previously implicated in passive impression formation.

8. Lindström B, Selbing I, Molapour T, Olsson A: **Racial bias shapes social reinforcement learning**. *Psychol Sci* 2014, **25**:711-719.

9. O'doherty JP, Hampton A, Kim H: **Model-based fMRI and its application to reward learning and decision making**. *Ann N Y Acad Sci* 2007, **1104**:35-53.

10. • Palminteri S, Wyart V, Koechlin E: **The importance of falsification in computational cognitive modeling**. *Trends Cogn Sci* 2017, **21**:425-433.

This paper proposes guidelines for testing which of several computational models best accounts for human behavior. The authors propose that researchers should test not only which model provides the best relative fit compared to other models (model comparison), but also whether each model succeds in reproducing features of the data (model falsification). This latter approach requires using simulations to identify how different models make qualitiatively different predictions about behavior.

11. Wood W, Rünger D: **Psychology of habit**. *Annu Rev Psychol* 2016, **67**:289-314 http://dx.doi.org/10.1146/annurev-psych-122414-033417.

12. Gläscher J, Daw ND, Dayan P, O'Doherty JP: **States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning**. *Neuron* 2010, **66**:585-595 http://dx.doi.org/10.1016/j.neuron.2010.04.016.

13. Thorndike EL: *Animal Intelligence: Experimental Studies*. New York: Macmillan; 1911.

14. Garrison J, Erdeniz B, Done J: **Prediction error in reinforcement learning: a meta-analysis of neuroimaging studies**. *Neurosci Biobehav Rev* 2013, **37**:1297-1310 http://dx.doi.org/10.1016/j.neubiorev.2013.03.023.

15. Jones RM, Somerville LH, Li J, Ruberry EJ, Libby V, Glover G, Voss HU, Ballon DJ, Casey BJ: **Behavioral and neural properties of social reinforcement learning**. *J Neurosci* 2011, **31**:13039-13045 http://dx.doi.org/10.1523/JNEUROSCI.2972-11.2011.

16. Lin A, Adolphs R, Rangel A: **Social and monetary reward learning engage overlapping neural substrates**. *Soc Cogn Affect Neurosci* 2011 http://dx.doi.org/10.1093/scan/nsr006.

17. Klucharev V, Hytonen K, Rijpkema M, Smidts A, Fernandez G: **Reinforcement learning signal predicts social conformity**. *Neuron* 2009, **61**:140-61151 http://dx.doi.org/10.1016/j.neuron.2008.11.027.

18. Nook EC, Zaki J: **Social norms shift behavioral and neural responses to foods**. *J Cogn Neurosci* 2015, **27**:1412-1426 http://dx.doi.org/10.1162/jocn_a_00795.

19. Kwak Y, Pearson J, Huettel SA: **Differential reward learning for self and others predicts self-reported altruism**. *PLoS One* 2014, **9** http://dx.doi.org/10.1371/journal.pone.0107621.

20. Lockwood PL, Apps MAJ, Valton V, Viding E, Roiser JP: **Neurocomputational mechanisms of prosocial learning and links to empathy**. *Proc Natl Acad Sci* 2016, **113**:9763-9768 http://dx.doi.org/10.1073/pnas.1603198113.

21. Sul S, Tobler PN, Hein G, Leiberg S, Jung D, Fehr E, Kim H: **Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality**. *Proc Natl Acad Sci* 2015, **112**201423895 http://dx.doi.org/10.1073/pnas.1423895112.

22. Lott AJ, Lott BE: **The role of reward in the formation of positive interpersonal attitudes**. In *Found. Interpers. Attract.* Edited by Huston TL. New York: Academic Press; 1974. p. 171.

23. Newcomb TM: **The prediction of interpersonal attraction**. *Am Psychol* 1956, **11**:575.

24. Clore GL, Byrne D: **A reinforcement-affect model of attraction**. In *Found. Interpers. Attract.*. Edited by Huston TL. New York: Academic Press; 1974:143-170.

25. Apps MAJ, Rushworth MFS, Chang SWC: **The anterior cingulate gyrus and social cognition: tracking the motivation of others**. *Neuron* 2016, **90**:692-707 http://dx.doi.org/10.1016/j.neuron.2016.04.018.

26. Ruff CC, Fehr E: **The neurobiology of rewards and values in social decision making**. *Nat Rev Neurosci* 2014, **15**:549-562 http://dx.doi.org/10.1038/nrn3776.

27. • FeldmanHall O, Dunsmoor JE, Kroes MCW, Lackovic S, Phelps EA: **Associative learning of social value in dynamic groups**. *Psychol Sci* 2017, **28**:1160-1170.

This study tested whether computational principles of Pavlovian learning can account for the development of complex social preferences. Participants learned about an altruistic giver who first shared money alone and later shared money in conjunction with a partner. Participants did not learn to associate the partner with value, consistent with blocking mechanisms predicted by associative learning models.

28. FeldmanHall O, Dunsmoor JE, Tompary A, Hunter LE, Todorov A, Phelps EA: **Stimulus generalization as a mechanism for learning to trust**. *Proc Natl Acad Sci U S A* 2018, **115**:E1690-E1697.

29. Amodio DM, Berg JJ: **Toward a multiple memory systems model of attitudes and social cognition**. *Psychol Inq* 2018, **29**:14-19.

30. Smith ER, DeCoster J: **Dual-process models in social and cognitive psychology: conceptual integration and links to underlying memory systems**. *Pers Soc Psychol Rev* 2000, **4**:108-131.

31. Ma N, Vandekerckhove M, Baetens K, Van Overwalle F, Seurinck R, Fias W: **Inconsistencies in spontaneous and intentional trait inferences**. *Soc Cogn Affect Neurosci* 2012, **7**:937-950 http://dx.doi.org/10.1093/scan/nsr064.

32. Mende-Siedlecki P, Cai Y, Todorov A: **The neural dynamics of updating person impressions**. *Soc Cogn Affect Neurosci* 2013, **8**:623-631 http://dx.doi.org/10.1093/scan/nss040.

33. Mende-Siedlecki P, Baron SG, Todorov A: **Diagnostic value underlies asymmetric updating of impressions in the morality**

and ability domains. *J Neurosci* 2013, **33**:19406-19415 http://dx.doi.org/10.1523/JNEUROSCI.2334-13.2013.

34. Mende-Siedlecki P, Todorov A: **Neural dissociations between meaningful and mere inconsistency in impression updating**. *Soc Cogn Affect Neurosci* 2016, **11**:1489-1500.

35. Boorman ED, O'Doherty JP, Adolphs R, Rangel A: **The behavioral and neural mechanisms underlying the tracking of expertise**. *Neuron* 2013, **80**:1558-1571 http://dx.doi.org/10.1016/j.neuron.2013.10.024.

36. Stanley DA: **Getting to know you: general and specific neural computations for learning about people**. *Soc Cogn Affect Neurosci* 2016, **11**:525-536 http://dx.doi.org/10.1093/scan/nsv145.

37. Chang LJ, Doll BB, van't Wout M, Frank MJ, Sanfey AG: **Seeing is believing: trustworthiness as a dynamic belief**. *Cogn Psychol* 2010, **61**:87-105.

38. Delgado MR, Frank RH, Phelps EA: **Perceptions of moral character modulate the neural systems of reward during the trust game**. *Nat Neurosci* 2005, **8**:1611-1618 http://dx.doi.org/10.1038/nn1575.

39. King-Casas B: **Getting to know you: reputation and trust in a two-person economic exchange**. *Science (80-)* 2005, **308**:78-83 http://dx.doi.org/10.1126/science.1108062.

40. Fareri DS, Chang LJ, Delgado MR: **Effects of direct social experience on trust decisions and neural reward circuitry**. *Front Neurosci* 2012, **6**:148.

41. Diaconescu AO, Mathys C, Weber LAE, Daunizeau J, Kasper L, Lomakina EI, Fehr E, Stephan KE: **Inferring on the intentions of others by hierarchical Bayesian learning**. *PLoS Comput Biol* 2014, **10**e1003810.

42. Behrens TEJ, Woolrich MW, Walton ME, Rushworth MFS: **Learning the value of information in an uncertain world**. *Nat Neurosci* 2007, **10**:1214.

43. Zhu L, Mathewson KE, Hsu M: **Dissociable neural representations of reinforcement and belief prediction errors underlie strategic learning**. *Proc Natl Acad Sci U S A* 2012, **109**:1419-1424 http://dx.doi.org/10.1073/pnas.1116783109.

44. Hampton AN, Bossaerts P, O'Doherty JP: **Neural correlates of mentalizing-related computations during strategic interactions in humans**. *Proc Natl Acad Sci U S A* 2008, **105**:6741-6746 http://dx.doi.org/10.1073/pnas.0711099105.

45. •• Hill CA, Suzuki S, Polania R, Moisa M, O'Doherty JP, Ruff CC: **A causal account of the brain network computations underlying strategic social behavior**. *Nat Neurosci* 2017, **20**:1142.
The authors combined computational modeling, fMRI, and TMS to understand neural bases of mentalizing during strategic decisions. TMS over right temporoparietal junction (rTPJ) disrupted the extent to which partucipants used a mentalizing-based strategy in a competitive game, as revealed in model fits and neural activity in rTPJ.

46. Coricelli G, Nagel R: **Neural correlates of depth of strategic reasoning in medial prefrontal cortex**. *Proc Natl Acad Sci U S A* 2009, **106**:9163-9168 http://dx.doi.org/10.1073/pnas.0807721106.

47. • Zaki J, Kallman S, Wimmer GE, Ochsner K, Shohamy D: **Social cognition as reinforcement learning: feedback modulates emotion inference**. *J Cogn Neurosci* 2016, **28**:1270-1282 http://dx.doi.org/10.1162/jocn_a_00978.
This study asked how people learn which cues to use when inferring another person's emotional state. The authors used a reinforcement learning model to characterize how perceivers adjust their emotion inferences in response to accuracy feedback, and found that trial-by-trial learning signals correlated with activation in both ventral striatum and TPJ.

48. Bandura A: *Social learning theory*. 1977.

49. Lindström B, Haaker J, Olsson A: **A common neural network differentially mediates direct and social fear learning**. *Neuroimage* 2018, **167**:121-129.

50. Dunne S, D'Souza A, O'Doherty JP: **The involvement of model-based but not model-free learning signals during observational reward learning in the absence of choice**. *J Neurophysiol* 2016, **115**:3195-3203.

51. Foerde K, Shohamy D: **Feedback timing modulates brain systems for learning in humans**. *J Neurosci* 2011, **31**:13157-13167 http://dx.doi.org/10.1523/JNEUROSCI.2701-11.2011.

52. Poldrack RA, Clark J, Paré-Blagoev EJ, Shohamy D, Creso Moyano J, Myers C, Gluck MA: **Interactive memory systems in the human brain**. *Nature* 2001, **414**:546-550 http://dx.doi.org/10.1038/35107080.

53. Berns GS, Capra CM, Moore S, Noussair C: **Neural mechanisms of the influence of popularity on adolescent ratings of music**. *Neuroimage* 2010, **49**:2687-2696.

54. Burke CJ, Tobler PN, Schultz W, Baddeley M: **Striatal BOLD response reflects the impact of herd information on financial decisions**. *Front Hum Neurosci* 2010, **4**:48.

55. Campbell-Meiklejohn DK, Bach DR, Roepstorff A, Dolan RJ, Frith CD: **How the opinion of others affects our valuation of objects**. *Curr Biol* 2010, **20**:1165-1170.

56. De Martino B, O'Doherty JP, Ray D, Bossaerts P, Camerer C: **In the mind of the market: theory of mind biases value computation during financial bubbles**. *Neuron* 2013, **79**:1222-1231.

57. Suzuki S, Adachi R, Dunne S, Bossaerts P, O'Doherty JP: **Neural mechanisms underlying human consensus decision-making**. *Neuron* 2015, **86**:591-602.

58. Crockett MJ: **How formal models can illuminate mechanisms of moral judgment and decision making**. *Curr Dir Psychol Sci* 2016, **25**:85-90.

59. Crockett MJ, Kurth-Nelson Z, Siegel JZ, Dayan P, Dolan RJ: **Harm to others outweighs harm to self in moral decision making**. *Proc Natl Acad Sci U S A* 2014, **111**:17320-17325.

60. Zaki J, López G, Mitchell JP: **Activity in ventromedial prefrontal cortex co-varies with revealed social preferences: evidence for person-invariant value**. *Soc Cogn Affect Neurosci* 2014, **9**:464-469 http://dx.doi.org/10.1093/scan/nst005.

61. •• Hutcherson CA, Bushong B, Rangel A: **A neurocomputational model of altruistic choice and its implications**. *Neuron* 2015, **87**:451-462.
This study applied a drift-diffusion model to altruistic decision-making. Activity in distinct neural regions correlated with distinct quantities predicted by the model. Modeling and neural data offered new interpretations of past neuroimaging findings, and suggested that generous choices may sometimes be errors.

62. Crockett MJ: **Models of morality**. *Trends Cogn Sci* 2013, **17**:363-366.

63. Cushman F: **Action, outcome, and value: a dual-system framework for morality**. *Pers Soc Psychol Rev* 2013, **17**:273-292.

64. Kleiman-Weiner M, Saxe R, Tenenbaum JB: **Learning a commonsense moral theory**. *Cognition* 2017, **167**:107-123.

65. Hackel LM, Zaki J: **Propagation of economic inequality through reciprocity and reputation**. *Psychol Sci* 2018, **29**:604-613.

66. FeldmanHall O, Otto AR, Phelps EA: **Learning moral values: Another's desire to punish enhances one's own punitive behavior**. *J Exp Psychol Gen* 2018, **147(8)**:1211-1224.

67. Wedekind C, Milinski M: **Cooperation through image scoring in humans**. *Science (80-)* 2000, **288**:850-852 http://dx.doi.org/10.1126/science.288.5467.850.

68. Nowak MA, Sigmund K: **Evolution of indirect reciprocity by image scoring**. *Nature* 1998, **393**:573-577.

69. Schiller D, Levy I, Niv Y, LeDoux JE, Phelps EA: **From fear to safety and back: reversal of fear in the human brain**. *J Neurosci* 2008, **28**:11517-11525.

70. Daw ND: **Trial-by-trial data analysis using computational models**. *Decis. Making, Affect. Learn. Atten. Perform. XXIII*. 2011.