Science
AAAS

The Computation of Social Behavior
Timothy E. J. Behrens, *et al.*
Science **324**, 1160 (2009);
DOI: 10.1126/science.1169694

Science (print ISSN 0036-8075; online ISSN 1095-9203) is published weekly, except the last week in December, by the American Association for the Advancement of Science, 1200 New York Avenue NW, Washington, DC 20005.

The Computation of Social Behavior

Timothy E. J. Behrens,* Laurence T. Hunt,* Matthew F. S. Rushworth*

Neuroscientists are beginning to advance explanations of social behavior in terms of underlying brain mechanisms. Two distinct networks of brain regions have come to the fore. The first involves brain regions that are concerned with learning about reward and reinforcement. These same reward-related brain areas also mediate preferences that are social in nature even when no direct reward is expected. The second network focuses on regions active when a person must make estimates of another person's intentions. However, it has been difficult to determine the precise roles of individual brain regions within these networks or how activities in the two networks relate to one another. Some recent studies of reward-guided behavior have described brain activity in terms of formal mathematical models; these models can be extended to describe mechanisms that underlie complex social exchange. Such a mathematical formalism defines explicit mechanistic hypotheses about internal computations underlying regional brain activity, provides a framework in which to relate different types of activity and understand their contributions to behavior, and prescribes strategies for performing experiments under strong control.

Social cognitive neuroscience attempts to identify and describe the brain areas and mechanisms that mediate social life. It is a field with a strange status. On the one hand, the proliferation of papers and conferences on the neural mechanisms of social cognition attest to a widespread excitement: Techniques for brain imaging now make it possible to investigate the biological basis of aspects of behavior that seem the most intrinsically human. On the other hand is the unusual skepticism with which such findings are sometimes greeted. Ultimately, behind this skepticism lies a concern that there may be something fundamental missing from our understanding of the neural mechanisms of social cognition that we take for granted when we study the neural mechanisms of other perceptual, cognitive, and motor processes.

If the most frequently studied aspects of social cognition are distinctively human, then is social cognitive neuroscience hampered by an absence of comparative models in other species? Like other perceptual, cognitive, and motor processes, does social cognition have a firm neuroanatomical basis? If so, can we understand its importance in terms of interconnections and interactions between brain areas? Does the complexity of social interactions prevent the investigation of neural mechanisms under controlled conditions? Are social cognitive neuroscientists therefore forced to rely on the fallacy of "reverse inference," misusing neural findings in an attempt to dissect cognitive processes (*1*)? Perhaps most important, is it feasible to describe the neural computations necessary to support social cognition in a way that allows precise and falsifiable predictions of our data in a framework that can be related across different studies? We contend that it is increasingly possible to understand social cognition in the context of an understanding of brain anatomy in human and other species, as well as in mathematical descriptions of behavior recorded in well-controlled experiments.

Clearly there are reasons for thinking that social cognition is an important brain function. According to the data gathered by advocates of the social brain hypothesis, the number and complexity of social interactions that an individual is likely to experience is a major determinant of interspecies differences in forebrain size, both generally in mammals and birds and, more specifically, when we focus on primates or even exclusively on hominoids (*2*). That social cognition should have such an impact is unsurprising when we consider the importance that social interactions have for individual survival and evolutionary fitness (*3*, *4*).

Several distinct approaches have been adopted when attempting to account for the brain basis of social cognition. One suggests that some brain areas have a uniquely social function (see below). Another approach, however, argues that an aggregation of simple, nonsocial processes will account for complex social behavior. Reward-guided behavior is known to depend on brain structures such as the orbitofrontal cortex (OFC) and amygdala (Fig. 1A). It is suggested that these structures might underlie the value associated with a particular person, just as they underlie values assigned to nonsocial stimuli (*5*). Other areas associated with reward and reinforcement, such as the ventral striatum and anterior cingulate cortex sulcus (ACCs), might also be expected to play a role. Some advocates of the social brain hypothesis appear to endorse such a view by arguing that social complexity is correlated with widespread differences in basic neuroanatomical features (such as brain size) rather than with changes limited to small, specialized brain regions.

Combining Formal Behavioral Models with Neurophysiological Data

It is clear that placing behavior in a social context does have a measurable effect on activity in brain regions associated with reward. For example, activity in ventral striatum that increases in receipt of monetary rewards also increases when subjects receive positive appraisals by their peers (Fig. 2A) (*6*).

In the domain of reward-guided behavior, our understanding of such signals has been transformed by recent attempts to provide an underlying mathematical formalism. Mathematical models that predict behavior bring a number of key advantages. Such models have different internal parameters that relate to different precise computations. By designing situations in which these model parameters fluctuate independently through trials, scientists can ask specific questions about neural activity (Fig. 1B) (*7*). These questions relate not to differences between different tasks but to the internal computations necessary to support a single task. In two different trials the stimuli and task might be identical, establishing strong control, but internal computations may differ (e.g., as a result of different previous experiences). Mathematical models make predictions of these internal computations that can be tested in neural data. By relating different parameters together in formal equations, they also predict precisely the effect that differences in neural activity should cause in behavior. In doing so, they obviate the possibility of reverse inference. Hypotheses in such studies are about not only brain regions, but also computational mechanisms.

A key example of this approach comes from reward-sensitive activity in midbrain dopamine neurons, which project to the ventral striatum. Dopaminergic activity does not, however, simply differentiate rewarded from unrewarded events. Rather, it codes a quantitative prediction of expected reward (derived from past experiences) and the quantitative deviation in observed reward from this prediction (*8*). This reward prediction error signal is an essential component of theoretical models of reinforcement learning (RL).

At their most simple, RL algorithms state that expectations of future reward (V_{t+1}) should be a function of current expectations (V_t) and their discrepancy from the actual outcome that is experienced—the prediction error (δ_t). More specifically, future expectations should be updated by the product of the prediction error and the learning rate (α_t) (*9*):

$$V_{t+1} = V_t + \alpha_t \delta_t \quad (1)$$

Many recent studies have used situations in which such parameters (V_t, α_t, δ_t, and other more complex ones) fluctuate independently from trial to trial and have simultaneously recorded neural activity electrophysiologically or with brain-imaging

Oxford Centre for Functional MRI of the Brain (FMRIB Centre), John Radcliffe Hospital, Oxford OX3 9DU, UK, and Department of Experimental Psychology, University of Oxford, Oxford OX1 3UD, UK.

*To whom correspondence should be addressed. E-mail: behrens@fmrib.ox.ac.uk (T.E.J.B.); lhunt@fmrib.ox.ac.uk (L.T.H.); matthew.rushworth@psy.ox.ac.uk (M.F.S.R.)

techniques. Trial-to-trial fluctuations in activity in particular cells or brain regions have been found to correlate with different model parameters, demonstrating dissociations and specializations in functional processing [for reviews, see (*10*, *11*)]. Despite addressing particular aspects of computation, these studies work in common or related mathematical frameworks. This mathematical framework therefore also serves as a conceptual framework for understanding the relationships between neural signals observed in different studies (*12*).

This approach is clearly well suited to modeling social interactions where an important aim is to obtain a reinforcer (usually money). In a two-person investment game, an investor is given money that he or she can choose to keep or invest with a trustee, with whom its value will triple. The trustee then decides how much money to return to the investor. Both players must consider their own actions and those of the opponent. Functional magnetic resonance imaging (fMRI) signal in the striatum of the trustee predicts the likelihood of the trustee to reciprocate investment (*13*). In early rounds of the game, this signal appears only after the investor has revealed his or her investment. In later rounds, however, when the trustee has experience of previous investments, the signal shifts to a period before the investment has been made (Fig. 2B). This temporal shift is reminiscent of the shift in dopaminergic prediction error activity from the time of a reward to the time of its predictive conditioned stimulus (*8*). Notably, although the implication of this time shift is that subjects make a neural prediction of opponent play, the striatal signal itself is predictive of the subject's own actions—whether to reciprocate a trusting investment.

Reward and Social Preferences

In social decision-making tasks, most people do not simply consider their own individual best interest. Instead, humans naturally perceive certain actions as rewarding or aversive because of their effect on other individuals. These "other-regarding" preferences can still be formalized within a framework in which people seek to maximize their expected benefit or utility by including terms that describe an aversion to inequality (β) (*14*). For example, if an event has outcome values V_i and V_j for players i and j, an inequality averse player i might only extract utility $U_i = V_i - \beta(V_i - V_j)$ from the event. In other words, it is argued, actors perceive the outcomes of interpersonal interactions from within a frame of reference that is tied to their own personal outcome, but the well-being of others impinges on the utility of this outcome. It is indeed the case that other-regarding preferences elicit neural responses that mirror those to standard rewarding or aversive situations (*14*). For example, acts of altruism thought to be personally rewarding elicit striatal activity, whether in the context of making charitable donations (*15*) or of punishing unfair players (Fig. 2, C and D) (*16*, *17*).

Other-regarding preferences thus serve to modify the value of a subject's own actions to account for his or her effect on other people. Like those of the trustee in the investment game (*13*) and the subjects under peer review (*6*), signals in the altruists' striatums are thus modified by social context, but still ultimately reflect the value of the social situation to the subject. It is unclear, therefore, whether such signals reflect the initial social cognitive events that lead to such a valuation or the consequent impact of social processing on valuation and behavior.

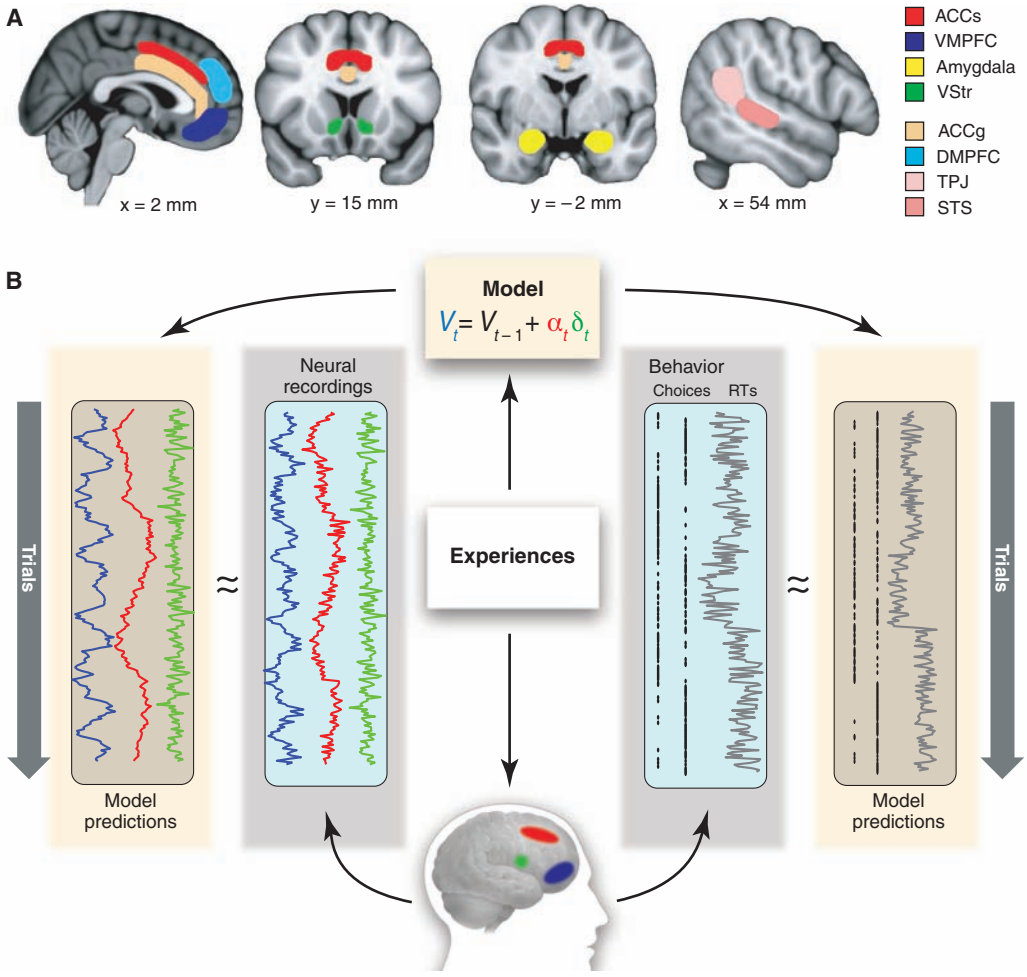

Fig. 1. (**A**) The functional neuroanatomy of social behavior. Primary colors denote brain regions activated by reward and valuation, frequently identified in studies of social interaction within the frame of reference of the subject's own actions: anterior cingulate cortex sulcus (ACCs), ventromedial prefrontal cortex (VMPFC), amygdala, and ventral striatum (VStr). Pastels denote brain regions activated by considering the intentions of another individual: anterior cingulate cortex gyrus (ACCg), dorsomedial prefrontal cortex (DMPFC), temporoparietal junction (TPJ), and superior temporal sulcus (STS). (**B**) Schematic of an approach that combines mathematical models of behavior with neural recordings. The model contains parameters that represent specific computations underlying behavior. As the subject/model undergoes different experiences, these parameters will fluctuate. The fluctuation in these parameters is used to find neural correlates of the specific underlying computations. Separately, the same parameter fluctuations come together to predict changes in behavior.

Functional Specializations for Social Behavior

It has been suggested that some brain regions (Fig. 1A) are involved only in social processing (*18*, *19*). The specificity of such areas' roles has been debated, and it is these areas' functions that have seemed particularly recalcitrant to formal description. A dorsomedial prefrontal region in the vicinity of the paracingulate sulcus is perhaps the most studied of these regions. This region is active during "theory of mind" (ToM) games played against other individuals, but not played against computers (*18*, *19*). Just what contribution the paracingulate region makes to the performance of such games has been harder to ascertain, but some distinct proposals have been made. Some accounts argue that its activation is a consequence of the joint attention that occurs between two individuals in ToM games (*19*). Other accounts ascribe the paracingulate cortex, or frequently co-activated

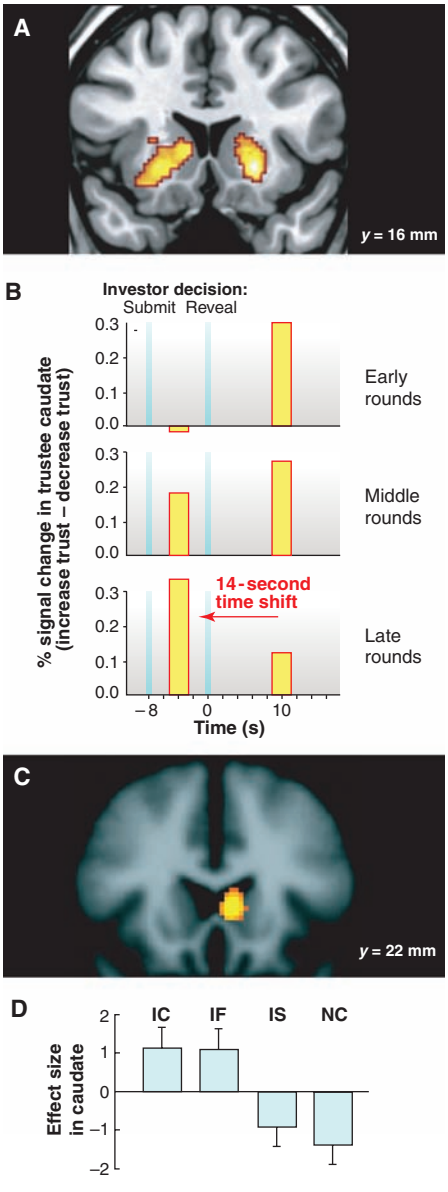

Fig. 2. Reward- and value-related striatal activity during social interactions parallels striatal activity in nonsocial tasks. (**A**) In the same subjects, a region of the caudate nucleus is activated by monetary rewards in a lottery task and is also activated by positive social appraisals. [Reprinted from *Neuron* (*6*) with permission from Elsevier] (**B**) In a reciprocal investment game, activity in the caudate nucleus of the trustee is greater on trials where the trustee's trust increases rather than decreases. In early rounds this signal is seen after the investor decision is revealed, but in later rounds the signal shifts to a time point before revelation, which suggests that the trustee builds a model of the investor's likely actions (*13*). (**C** and **D**) Altruistic punishment of unfair partner behavior in an economic exchange activates the punisher's caudate nucleus (C) in conditions where the punishment costs the partner money [IC and IF in (D)] but not in conditions where the punishment is symbolic (IS) or randomly selected (NC) (*16*).

regions such as the posterior superior temporal sulcus (STS) and temporoparietal junction (TPJ), with roles in metacognition—thinking about the intentions of another person (*18*–*20*). Although it is certainly the case that one's own simple motor intentions do not activate the same regions (*21*, *22*), it has been argued that it is difficult to test with precision the function that is accomplished by these regions, because of the complex set of different mental and neural processes that may be recruited in ToM situations. However, computational approaches (see below) have recently been used to generate quantitative predictions about how activity should change in these brain areas when subjects estimate, and revise their estimations of, another's intentions.

A distinct strand of research has emphasized the importance of parts of the cingulate cortex for social processing. fMRI studies have showed signal increases in two divisions of the cingulate cortex, one in the posterior cingulate cortex adjacent to retrosplenial areas (*19*, *22*) and one in the anterior cingulate cortex gyrus (ACCg) (*23*), when people are engaged in processes that are prerequisite for social cognition, such as the consideration of intentions or emotions. Again, accounts of such activity patterns may be sharpened by computational models that make quantitative predictions about the value of social information to an experimental participant.

Frames of Reference in Social Decision Making

It initially appears difficult to reconcile the emphasis on RL and on brain areas such as the striatum with the quite distinct brain areas revealed by investigations of ToM. It becomes less surprising when one remembers that social preference studies purposively avoid iterative game settings precisely because they want to avoid confounding a social preference for another's payoff with beliefs about another's intentions.

One potential consilience of ToM and reinforcement-based approaches comes from considering the different frames of reference they use in interrogating the data. As discussed above, social preference and game theory tasks often investigate signals in the frame of reference of the subjects' own actions or rewards. In ToM tasks, the analysis identifies areas that are more active when another individual's intentions must be decoded. Evidence for this idea comes from a study that combines an investigation of brain activity in an action observation task and subjective reports of a key social preference: altruism (*24*). Unlike the effects of altruistic behavior on the striatum (*15*), when the frame of reference adopted is not one of value or reward but rather one of intention, it is fMRI activity in a ToM region—the STS—that predicts a subject's altruistic tendencies (*24*).

It is important to recognize the quite distinct frame of reference that is adopted in these different approaches to social decision-making. Very recently, computational accounts have been extended to describe not just how value representations are learned during interpersonal interactions, but also the way in which representations of others' intentions evolve in such settings.

Social Computations

There is evidence that predictions in the frame of reference of an opponent exist at the level of single neurons. In a matching-pennies game, a trial is rewarded if the subject makes the same choice as an opponent. When monkeys played against a computer opponent, neurons were found whose activity was dependent on the monkeys' histories of choices and rewards (*25*, *26*). Intriguingly, however, some neural responses were also dependent on current and previous choices of the computer opponent (*26*), enabling a prediction of the computer's next play. Important questions remain. First, it is unclear whether the monkeys believed they were interacting with an intentional agent, and hence whether activity would occur similarly in social interactions. Second, although it is clear that such neurons are encountered in the dorsolateral prefrontal cortex, little is known about their wider distribution, and recordings have not been made in ToM-related brain areas. Nonetheless, by explicitly decoupling monkey behavior, opponent behavior, and predicted or experienced rewards, this experiment demonstrated separable neuronal activity in all three frames of reference.

Such a decoupling has also been used by two recent human fMRI studies that scanned the whole brain and aimed to dissociate two sorts of neural computations: those associated with predicting the behavior of another individual, and those that determine the effect of these predictions on outcome valuation and hence behavior (*27*, *28*). These studies test whether key concepts that underlie RL may generalize to the domains of social reputation learning (*27*) and mentalizing (*28*). In both studies, subjects were asked to play a game that required them to track and predict the behavior of other individuals (confederates) to optimize their own behavior. Formal computational models, based on principles from RL, were built that tracked this information probabilistically, which allowed for fMRI data to be interrogated for correlates of these models' parameters. In both cases, the use of formal mathematical modeling ensured that key computational variables relating to confederate behavior would fluctuate trial-by-trial, and that these variables would be decoupled from variables relating to reward processing. In both cases, because the mathematical models contained separate parameters relating to expectations of reward and confederate behavior, activity could be assessed separately in each frame of reference. Furthermore, despite considerable differences in the two tasks, the common mathematical framework allows neural signals in the two studies to be compared directly.

Prediction Errors on Confederate Behavior

In one case (*27*), subjects played a game that required them to learn in parallel about the likely

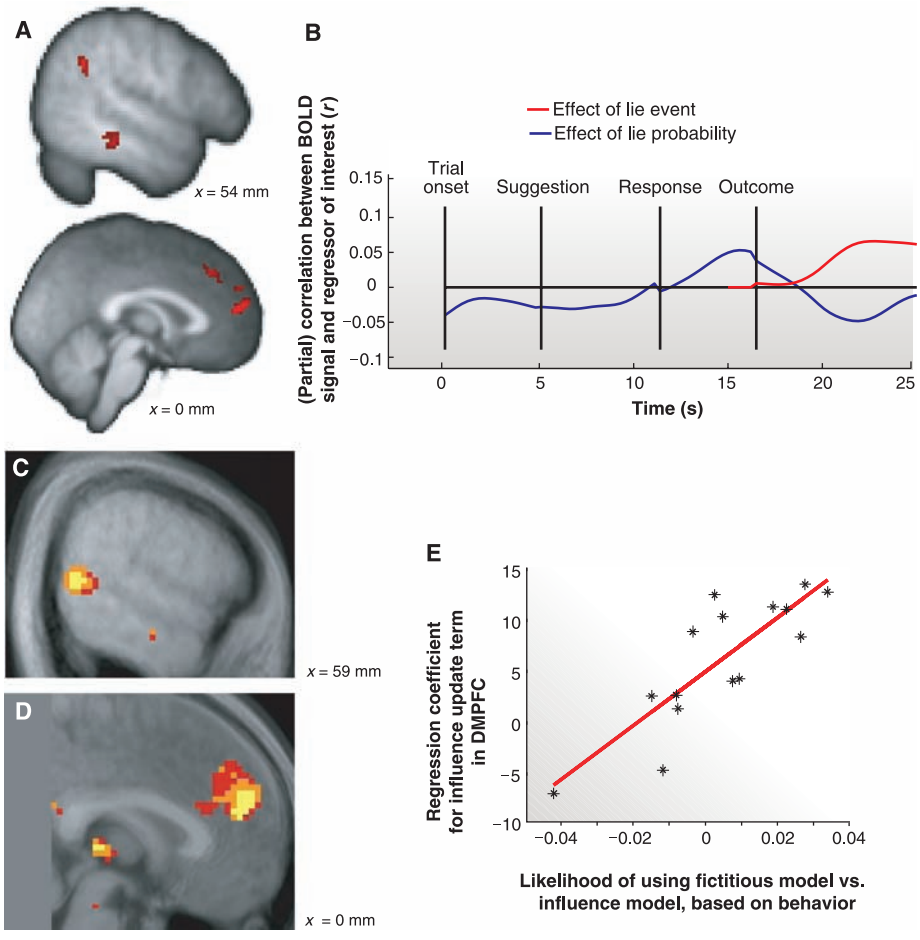

Fig. 3. Prediction errors on the intentions of a social partner's behavior activate theory of mind regions. (**A**) The STS/TPJ, middle temporal gyrus, and DMPFC all correlate with prediction error (Δ_t) on the probability of a confederate lie during a social reputation–building task. (**B**) Time course of activity shows all three components of a prediction error signal: positive correlate of lie probability before revelation of partner behavior ("outcome"), negative correlate of lie probability after, and positive correlate of lie event after. [Adapted by permission from Macmillan Publishers Ltd., *Nature* (*27*), copyright 2008] (**C**) During a "work-or-shirk" game, the STS signals the influence update of the subject's current action on the likely future behavior of the other player. (**D** and **E**) In the DMPFC (D), this signal correlates with the likelihood that the subject was using this "influence" model versus a simpler, "fictitious learning" model (E) (*28*). [Copyright 2008, National Academy of Sciences, U.S.A.]

location of a reward, and about the motives of a confederate advising them on their next choice. After each outcome, subjects were assumed to update the probability of a reward on a choice option (for example, the green rather than the blue card) using RL mechanisms outlined above: $V_{t+1} = V_t + \alpha_t\delta_t$. The same outcome, however, enabled subjects to independently update a running estimate of the probability of unfaithful confederate advice: $L_{t+1} = L_t + \beta_t\Delta_t$, where L_t represents the probability the confederate will lie at the current trial, and Δ_t and β_t represent a prediction error and learning rate on this probability, respectively. Two regions formerly isolated by ToM studies—paracingulate and STS/TPJ—performed a computational role that was directly analogous to dopaminergic activity in the reward domain. Activity first correlated with the probability that the confederate would lie. Subsequently, when the outcome was revealed, activity correlated with the quantitative and signed prediction error on confederate behavior (Fig. 3, A and B). Unlike in the dopaminergic system, the predictions and prediction errors at stake did not concern the scalar value of actions; instead, they concerned the truth of communicative intentions.

In another case (*28*), subjects played a repetitive Inspector game in which workers decide whether to work or shirk at each trial and Inspectors decide whether or not to inspect. Social interactions in such a game are complex. The worker should only work if he or she believes the inspector will inspect. The inspector should only embark on costly inspections if the worker is likely to shirk. In such a task, it is possible for subjects not only to track the intentions of an opponent [as was optimal in (*26*)], but also to try to second-guess the influence of their own actions on opponent behavior. This influence is also amenable to prediction error learning. Here the prediction error represents a deviation in subjects' own behavior (Q) from what they believe the opponent is predicting they will do (q^{**}):

$$\Delta_t = (Q_t - q_t^{**}) \quad (2)$$

fMRI correlates of such a prediction error are observed in the STS (Fig. 3C), not before the subject chooses, but when players witness the other's choice and must update their predictions of future opponent plays. In a paracingulate region, similar activations are witnessed, but predominantly in subjects where a reliance on this influence can be measured in their behavior (Fig. 3, D and E).

Parallel Processing of Reward and Intentional Reference Frames in the Cingulate Cortex

Evidence that key computational parameters, such as prediction errors, can be found in the social domain invites immediate parallels to be drawn between mechanisms underlying social and nonsocial behaviors. Such comparisons may serve to clarify functional dissociations between anatomical regions. For example, several quite distinct strands of evidence suggest an anatomical subdivision within the ACC between social and nonsocial behavior. Social tasks tend to activate the more ventral ACCg (Fig. 1B), whereas reward-guided behavior tends to activate a more dorsal region in the ACCs (Fig. 1A). Such a dissociation may depend on the connection patterns of these regions. In the monkey the ACCs is strongly interconnected with premotor and motor cortex, whereas the ACCg is connected with brain regions concerned with the processing of emotion, facial expression, and biological motion such as the hypothalamus and STS (*29*). Recent studies reveal similar gradients of connectivity across the human cingulate cortex (*23*) (Fig. 4A). Functionally, both the human and macaque ACCs appear to play a role in determining the value of new pieces of information for guiding future behavior (*10*, *30*–*32*). There is evidence suggesting a similar role for the ACCg in the social behavior of both species (*33*, *34*). For instance, healthy male macaques will forego food rewards to view images of certain high-status conspecifics (*35*). Lesions of the ACCg, but not the ACCs, diminish the value assigned to such social information in comparison to food rewards (*33*) (Fig. 4B). Finally, manipulations of human behavior also implicate the ACCg in social cognition. In human subjects, nasal administration of the neuropeptide oxytocin increases the trust that subjects place in confederates in social games (*36*). In early rounds of these games, the ACCg signal is normally greater than in a nonsocial decision-making task; under oxytocin, however, this increased signal disappears (*34*).

This evidence suggests not only that a functional dissociation exists between the ACCg and

ACCs for social and nonsocial behavior, but also that the two regions might play a similar role in the two domains. However, the evidence remains circumstantial, as it was gathered using several different approaches in very different experimental situations. Using formal computational models, however, it was possible to compare directly the computations performed in these two ACC subregions, in the same group of subjects performing the same task (*27*).

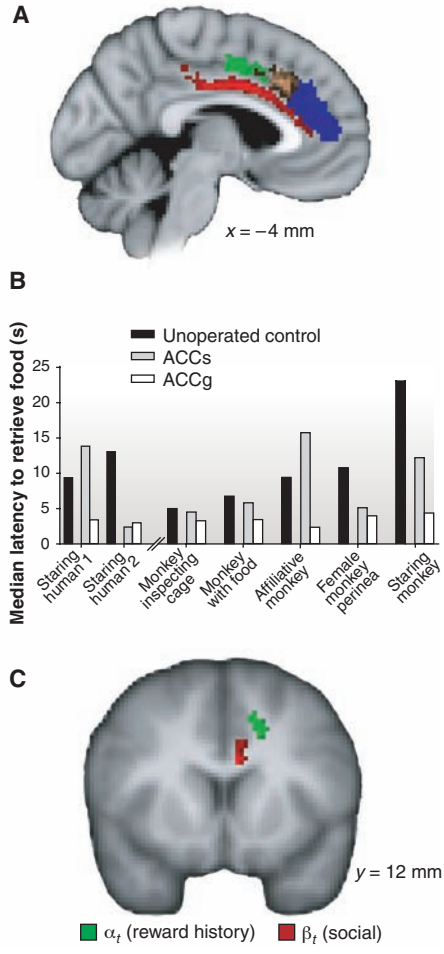

Fig. 4. Converging evidence from anatomy, lesion studies, and computational modeling of fMRI data for a dorsoventral dissociation in ACC. (**A**) Parcellation of cingulate cortex reveals a cluster in ACCg, and three in ACCs, with distinct connectivity patterns to other brain regions (*23*). (**B**) Monkeys are presented simultaneously with food and socially salient stimuli. In control monkeys, the most salient social stimuli (abscissa) induce the longest delays (ordinate) before the food is taken. This effect is abolished in monkeys with lesions to the ACCg but not to the ACCs (*33*). (**C**) Parallel learning rates in a reinforcement learning model for social and reward-based information are signaled in the ACCg and ACCs, respectively; the ACCg signal correlates with degree to which subjects use social information in the task, whereas the ACCs signal correlates with degree to which subjects use reward information. [Adapted by permission from Macmillan Publishers Ltd., *Nature* (*27*), copyright 2008]

During probabilistic learning, the value of a new piece of information can be defined formally. It dictates the instantaneous learning rate (α_t) (Eq. 1) that is used to weigh the current prediction error. In the parallel learning situation used in (*27*), independent fluctuations in the two learning rates for reward-guided (α_t) and social (β_t) learning could be seen in the fMRI data. As previously demonstrated (*30*), the reward-guided learning rate predicted fMRI fluctuations in the ACCs, particularly in individuals who would be more influenced by reward information. However, the ACCg signal reflected the social learning rate, particularly in individuals whose behavior was likely to be guided by confederate advice. The ACCs and ACCg therefore encoded the exact same computational parameter (the instantaneous learning rate) at the exact same time, but in two distinct frames of reference (Fig. 4C).

Notably, computational models of confederate behavior can also make predictions in the more traditional frame of reference—one's own actions. When analyzed in this frame of reference (*27*, *28*), activations closely matched those found in nonsocial studies (*10*, *11*). At the time of feedback, both studies identified a reward prediction error (δ_t in Eq. 1) of the subject's chosen action in the ventral striatum. At the time of action selection, a correlate of the expected value of the subject's chosen action (V_t in Eq. 1) was found by both studies in a ventromedial portion of the prefrontal cortex. Furthermore, despite being analyzed in the frame of reference of reward, in both studies these signals were best explained by models that accounted for the influence of confederate play on the subject's valuations.

Conclusions

There has been an unacknowledged tension between different neural accounts of social behavior. Some have focused on a number of brain regions with a general role in reinforcement processing (*5*). Other accounts have emphasized the importance of a circumscribed circuit concerned with the representations of other's beliefs and intentions (*18*, *19*). It is clear that it is necessary to draw on both traditions. RL-based approaches in conjunction with paradigms drawn from game theory have begun to describe the computations performed by reinforcement-related brain regions, such as the ventral striatum, during the course of social interaction as one person attempts to predict the behavior of another. These models focus on predictions and prediction errors that are tied to the frame of reference of the actor and consider the scalar value that the actor expects from the interaction. Increasingly, however, a similar formalism is being translated to the modeling of beliefs and the joint relationships between one's own actions and intentions and those of another person. Rather than focusing just on scalar value, the emphasis is on predictions about the truth of intentions or the truth of the relationship between an actor and another person. By considering such complex behaviors and signals in a formal fashion (*37*), relationships can be established between signals seen in different experimental situations. It is hoped that these formal relationships will reflect mechanistic properties of neural activity that will generalize across many kinds of social interactions.

References and Notes

1. R. A. Poldrack, *Trends Cognit. Sci.* **10**, 59 (2006).
2. R. I. Dunbar, S. Shultz, *Science* **317**, 1344 (2007).
3. K. E. Holekamp, S. T. Sakai, B. L. Lundrigan, *Philos. Trans. R. Soc. London Ser. B* **362**, 523 (2007).
4. J. B. Silk, *Science* **317**, 1347 (2007).
5. E. T. Rolls, *The Brain and Emotion* (Oxford Univ. Press, Oxford, 1999).
6. K. Izuma, D. N. Saito, N. Sadato, *Neuron* **58**, 284 (2008).
7. G. Corrado, K. Doya, *J. Neurosci.* **27**, 8178 (2007).
8. W. Schultz, P. Dayan, P. R. Montague, *Science* **275**, 1593 (1997).
9. R. A. Rescorla, A. R. Wagner, in *Classical Conditioning II: Current Research and Theory*, A. H. Black, W. F. Prokasy, Eds. (Appleton-Century-Crofts, New York, 1972), pp. 64–99.
10. M. F. Rushworth, T. E. Behrens, *Nat. Neurosci.* **11**, 389 (2008).
11. M. L. Platt, S. A. Huettel, *Nat. Neurosci.* **11**, 398 (2008).
12. N. D. Daw, K. Doya, *Curr. Opin. Neurobiol.* **16**, 199 (2006).
13. B. King-Casas *et al.*, *Science* **308**, 78 (2005).
14. E. Fehr, C. F. Camerer, *Trends Cognit. Sci.* **11**, 419 (2007).
15. J. Moll *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 15623 (2006).
16. D. J. de Quervain *et al.*, *Science* **305**, 1254 (2004).
17. T. Singer *et al.*, *Nature* **439**, 466 (2006).
18. D. M. Amodio, C. D. Frith, *Nat. Rev. Neurosci.* **7**, 268 (2006).
19. R. Saxe, *Curr. Opin. Neurobiol.* **16**, 235 (2006).
20. F. Castelli, C. Frith, F. Happe, U. Frith, *Brain* **125**, 1839 (2002).
21. N. Ramnani, R. C. Miall, *Nat. Neurosci.* **7**, 85 (2004).
22. H. E. den Ouden, U. Frith, C. Frith, S. J. Blakemore, *Neuroimage* **28**, 787 (2005).
23. M. Beckmann, H. Johansen-Berg, M. F. Rushworth, *J. Neurosci.* **29**, 1175 (2009).
24. D. Tankersley, C. J. Stowe, S. A. Huettel, *Nat. Neurosci.* **10**, 150 (2007).
25. D. J. Barraclough, M. L. Conroy, D. Lee, *Nat. Neurosci.* **7**, 404 (2004).
26. H. Seo, D. Lee, *Philos. Trans. R. Soc. London Ser. B* **363**, 3845 (2008).
27. T. E. Behrens, L. T. Hunt, M. W. Woolrich, M. F. Rushworth, *Nature* **456**, 245 (2008).
28. A. N. Hampton, P. Bossaerts, J. P. O'Doherty, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 6741 (2008).
29. G. W. Van Hoesen, R. J. Morecraft, B. A. Vogt, in *Neurobiology of Cingulate Cortex and Limbic Thalamus*, B. A. Vogt, M. Gabriel, Eds. (Birkhäuser, Boston, 1993).
30. T. E. Behrens, M. W. Woolrich, M. E. Walton, M. F. Rushworth, *Nat. Neurosci.* **10**, 1214 (2007).
31. M. Matsumoto, K. Matsumoto, H. Abe, K. Tanaka, *Nat. Neurosci.* **10**, 647 (2007).
32. S. W. Kennerley, M. E. Walton, T. E. Behrens, M. J. Buckley, M. F. Rushworth, *Nat. Neurosci.* **9**, 940 (2006).
33. P. H. Rudebeck, M. J. Buckley, M. E. Walton, M. F. Rushworth, *Science* **313**, 1310 (2006).
34. T. Baumgartner, M. Heinrichs, A. Vonlanthen, U. Fischbacher, E. Fehr, *Neuron* **58**, 639 (2008).
35. R. O. Deaner, A. V. Khera, M. L. Platt, *Curr. Biol.* **15**, 543 (2005).
36. M. Kosfeld, M. Heinrichs, P. J. Zak, U. Fischbacher, E. Fehr, *Nature* **435**, 673 (2005).
37. W. Yoshida, R. J. Dolan, K. J. Friston, *PLoS Comput. Biol.* **4**, e1000254 (2008).

38. Supported by the UK Medical Research Council (T.E.J.B., M.F.S.R.) and the Wellcome Trust (L.T.H.).

10.1126/science.1169694