Neuron
Perspective

Neuroeconomic Approaches to Mental Disorders

Kenneth T. Kishida,[1] Brooks King-Casas,[1,2] and P. Read Montague[1,2,*]
[1]Department of Neuroscience and Computational Psychiatry Unit
[2]Menninger Department of Psychiatry and Behavioral Sciences
Baylor College of Medicine, Houston, TX 77030, USA
*Correspondence: read@bcm.edu
DOI 10.1016/j.neuron.2010.07.021

The pervasiveness of decision-making in every area of human endeavor highlights the importance of understanding choice mechanisms and their detailed relationship to underlying neurobiological function. This review surveys the recent and productive application of game-theoretic probes (economic games) to mental disorders. Such games typically possess concrete concepts of optimal play, thus providing quantitative ways to track when subjects' choices match or deviate from optimal. This feature equips economic games with natural classes of control signals that should guide learning and choice in the agents that play them. These signals and their underlying physical correlates in the brain are now being used to generate objective biomarkers that may prove useful for exposing and understanding the neurogenetic basis of normal and pathological human cognition. Thus, game-theoretic probes represent some of the first steps toward producing computationally principled, objective measures of cognitive function and dysfunction useful for the diagnosis, treatment, and understanding of mental disorders.

The Biological Value of Gambling

The idea of a game typically rouses feelings of amusement and play. Over the last several hundred years, games have taken on a more cerebral bearing and are now an accepted approach to understanding important features of strategic interactions in the broadest sense. Games and the game theories that prescribe their optimal solutions have infiltrated many areas of application. The list is long and growing, including war games between groups of humans (Kahn, 1960), predator-prey games (Maynard-Smith, 1982), strategic interactions between mobile organisms and their natural environments, economic exchanges between individuals and institutions (Niederle and Roth, 2003; Roth, 2008), and so on. In the economics literature, a game is often depicted as a *decision problem with structure* so that one's payoffs can depend on one's own choices and some other input. For an organism, the biggest gambling game is investing its limited energetic resources into pursuing some prey or food source in the presence of uncertainty about the likely yields from such efforts. Consequently, there is biological survival value, and thus selective pressure, in being a good gambler in the real world, where too many bad gambles result in the ultimate loss—death.

Ironically, gambling games represent one of the most insidious applications of games to real-world biology. We say insidious because gambling and related addictions represent a huge medical, social, and fiscal problem. Despite these problematic outcomes, gambling games are now being used to probe the neural and cognitive substrates of decision-making. This usage makes sense. Games of chance are literally as old as recorded history and provide excellent probes of many important features of motivated choice. They require that an agent frame a situation (understand the goals and outcomes of a game), estimate and value possible outcomes and trajectories leading to outcomes, and make choices conditioned on these valuations. It is no secret that gambling games work. Games of chance found in casinos routinely extract enormous amounts of valuable resources (typically cash) from otherwise healthy human beings. Why is this? The simple answer is that the structure of gambling games is not arbitrary—they coevolved over thousands of years around human nervous systems. Therefore, gambling games have design features that exploit exactly the frailties in human decision-making systems while also maintaining the human desire to play them. These games are now being used as sensitive probes of human valuation machinery. In humans and nonhuman primates, such probes are now producing new insights when paired with measures of neural function (Camerer, 2003; Glimcher and Rustichini, 2004; Lee, 2005; Camerer and Fehr, 2006; Pessiglione et al., 2006; Lohrenz et al., 2007; Rangel et al., 2008; Montague and Assad, 2008; DeMartino et al., 2010).

Gambling Games and the Neural Processing of Uncertainty

As described above, gambling games have a particularly potent impact on human behavior. They have evolved in the context of humans to exploit limitations in our capacity to estimate probabilities. These limitations take many forms, including our well-known penchant for overrepresenting rare events, our rapid overgeneralization on the basis of scant data, and numerous value-illusions to which humans are subject (e.g., Tversky and Kahneman, 1974, 1982; Kahneman and Tversky, 2000). For this discussion, we focus on probabilistic games where behavioral options come equipped with payoff probabilities and returns, and players may know these well or not at all. The strength of using the games is that one can calculate ahead of time the likelihoods of payoffs and the risks involved, and, as the game progresses, one can update these computations in a principled way. Any problems in estimating the value of an option, the risk, and the likely payoffs will be mirrored by estimable changes in returns—usually losses. And one would expect such problems in the presence of brain damage,

addiction(s), and psychopathologies, which highlights the value of using gambling game probes in human populations.

One excellent example of using a gambling game to probe risk-dependent choice and its connection to neural systems was carried out by Hsu et al. (2005). One goal of this experiment was to uncover neural differences in the brain's response to two kinds of uncertainty—risk and ambiguity. In economic circles, these terms are distinguished by the confidence in the probabilities assigned to outcomes. Risk is the probability (judged by any means) of possible outcomes and their associated payoffs—in practice, this often means that such probabilities are known exactly for the outcomes, that is, no error bars on the probabilities. Ambiguity is also such a probability assignment, but in the presence of little, bad, or no evidence for having confidence in the assignment. There is some sloppiness here regarding the use of these terms, but ambiguity implies something like the "biggest error bars possible" or "terribly large error bars" on the probabilities while the source of this varying lack of confidence doesn't matter. In real-world experiments, humans don't like ambiguity, and their nervous systems will indeed take account of the confidence in an estimate. Hsu et al. examined limiting cases for how a human nervous system deals with uncertainty and then applied the same experiment to subjects with lesions in the orbitofrontal cortex. The main finding that emerged was that a frank lesion in orbitofrontal cortex correlated with a lack of sensitivity to risk and ambiguity. This was not true of a matched "control lesion group" who possessed lesions elsewhere in the brain, yet still displayed sensitivity to both risk and ambiguity as probed by their task. Huettel and colleagues used an even more extreme notion of ambiguity where the probabilities are unknown and showed that this particular case may involve dedicated neural systems (Huettel et al., 2006; also see Tom et al., 2007). In a remarkable two-subject experiment, DeMartino et al. (2010) used monetary gambles to show that subjects with focal bilateral amygdala damage had a dramatic reduction in loss aversion compared to matched controls, a result consistent with the amygdala's role in response inhibition to fear-inducing stimuli (e.g., see Delgado et al., 2008).

These findings collectively suggest that loss of identified brain tissue correlates with a complete loss or at least dramatic diminution in sensitivity to uncertainty. Why review such evidence in a piece relating neuroeconomic probes to mental disorders? Let's answer this by considering the problem the other way around. Imagine that we profiled a large population of subjects on risk and ambiguity sensitivity experiments, found a subpopulation with diminished sensitivities along these dimensions, and chased this maneuver with brain imaging (structural and function) experiments. The economic probe becomes a way to uncover possible problems in orbitofrontal cortex exposed by the demands of a risk-dependent task like that illustrated by Hsu et al. (2005). We have the example from the lesion patients that loss of orbitofrontal tissues diminishes sensitivity to risk (e.g., Clark et al., 2008), but subtler changes in this region could express the same behavioral changes seen with injuries or changes in gene expression in the area. Screening on sensitivity to risk and ambiguity also has important practical uses for mental disorders where impulsivity, valuation, and decision-making can all be concurrently perturbed. This is also an important area to pursue because of the prevalence of frontal head injuries in sporting events, war zones, and recreational activity that, while not resulting in structurally detectable damage, could well be detectable by differences in functional imaging signatures in this or other regions.

Biomarkers Exposed by Gambling Games

The relatively blunt demonstration of how altered function in select brain regions (i.e., loss of function due to a lesion) alters computations of valuation and decision-making behavior preludes the role gambling games will have in determining biomarkers of abnormal decision-making as expressed in a range of mental disorders. Recent efforts to characterize biomarkers of mental disorders using gambling games include examples in drug addiction and schizophrenia.

Recently, Chiu et al. assessed sated and unsated nicotine addicts compared to nonsmoker controls using a sequential investment task and fMRI (Chiu et al., 2008b; Lohrenz et al., 2007). The sequential investment task (Lohrenz et al., 2007) was designed around current models of dopaminergic function and the notion that model-based parameters could be used to extract physiological correlates—biomarkers—of rational control signals. This model-based approach yielded insight into quantifiable parameters that predict behavior, the neural basis of control signals underlying choice behavior, and how brains addicted to nicotine express altered behavioral responsivity to these control signals. Chiu and colleagues demonstrate that control subjects' subsequent choice in the sequential game is best predicted by a normative control signal, the fictive error, and show that the caudate nucleus generates a signal consistent with computation of this fictive error. The fictive error as expressed in this game is a counterfactual signal about "what could have been," if the subject had made a different decision (invested more or less) and is parameterized by the change in market price and the subject's investment level (see Camerer and Ho, [1999] for earlier approaches to reinforcement learning models that incorporate parameters for counterfactual signals analogous to the fictive error). Smokers demonstrated the fictive error signal in the same brain region, but their behavior was no longer predicted by this computation—suggesting that computation of the fictive error term was intact in the smokers' brain, but was disconnected from influencing their behavior. This computational term is consistent with the gains smokers knowingly forego (money savings, increased quality of health, and length of life) in order to continue their habit.

Efforts to estimate parameters related to impulsive behaviors often expressed in drug addicts have also utilized simple gambling games aimed at determining temporal discounting behavior. Temporal discounting tasks consist of choices like "choose one of two options," option A: money now (say $2) or option B: more money ($10) at some later time, t. The cost associated with waiting to acquire some dollar amount later (temporal delay) discounts the value of the later option and is known as temporal discounting behavior. This kind of gamble is ecologically relevant to everyday human decisions: "do I fill up my gas tank now at some cost, or do I wait and risk not finding gas later at a better price, or worse yet, running out of gas before finding another station." Bickel and colleagues demonstrated that

a hyperbolic discounting model accurately characterizes and differentiates choice behavior in humans addicted to nicotine compared to nonaddicts during a temporal discounting task (Bickel et al., 2008). Nonsmokers expressed hyperbolic discounting behavior, as did smokers; however, the slope of the discount function for smokers is significantly steeper. They discount future gains at a faster rate than nonsmokers. The slope of the discount function is suggestive of a parameter for impulsivity that ought to show natural variability in the human population.

Paulus and colleagues have used neuroimaging and simple gambling tasks to assess choice behavior in methamphetamine addicts. In sum, the tasks used were variations of simple gambling games: (1) two-choice gamble with variable reward probabilities (Paulus et al., 2003b; Vollenweider et al., 2005; Paulus et al., 2005, 2008), (2) a risky temporal delay task (Leland and Paulus, 2005), and (3) a card prediction game with parameterized levels of certainty and probability of success of choice outcomes (Leland et al., 2006; Critchley et al., 2001). These studies took advantage of a number of simple gambles and models of subjects' behavior given the rational constraints of the tasks to determine measures that distinguished the stimulant users from the control groups. These experiments demonstrate that methamphetamine addicts can be distinguished from controls via altered brain activation and behavioral patterns that are consistent with poor decision-making in these subjects, including increased risk taking (Leland and Paulus, 2005), increased impulsivity (Leland et al., 2006; Paulus et al., 2008), and altered brain and behavioral responses to errors (Paulus et al., 2003b; Vollenweider et al., 2005; Paulus et al., 2008). Notably, in one study the parameters extracted from behavioral patterns expressed during a simple two-choice gamble with and without uncertainty and the associated brain responses measured with fMRI could be used to predict relapse in a cohort of treatment-seeking methamphetamine addicts (Paulus et al., 2005).

Drug addicts by definition make poor decisions (continued drug abuse in the face of adverse consequences), and the fact that gambling games can identify quantifiable behavioral characteristics and neuroimaging-based biomarkers that distinguish users from nonusers is promising. Evidence from early work suggests that these paradigms will also be useful for identifying biomarkers in other mental disorders, including schizophrenia (Paulus et al., 2003a), neuroticism outcomes (Feinstein et al., 2006; Paulus et al., 2003c), high-trait anxiety (Paulus et al., 2004), panic disorder (Ludewig et al., 2003), bipolar disorder (Minassian et al., 2004), and autism (Minassian et al., 2007). In these studies, a number of parameters are determined from choice behavior on simple gambling games. Associated alterations in brain responses are demonstrated for subjects with schizophrenia (Paulus et al., 2003a), neuroticism (Feinstein et al., 2006; Paulus et al., 2003c), and high-trait anxiety (Paulus et al., 2004). Behavioral characteristics derived from model-based analyses of sequences of choice behavior also differentiate control participants from those with schizophrenia (Paulus et al., 1999, 1996), panic disorder (Ludewig et al., 2003), bipolar disorder (Minassian et al., 2004), and autism (Minassian et al., 2007). The specificity of these behavioral parameters and the associated brain responses has yet to be determined, and to call these signals true biomarkers is still premature. However, these studies suggest that biomarkers for mental disorders may be determined using human neuroimaging paired with game-theoretic probes, which possess critical features (i.e., equilibrium solutions and best responses) for more sophisticated model-based approaches.

Equilibrium Solutions and Best Responses

One of the more important features of game-theoretic probes is the solution concept for the game, that is, the way that rational self-interested players should play the game in order to maximize their own returns. Solution concepts for game-theoretical probes are valuable because they can be used to guess the kinds of control or learning signals that an organism would need to generate in order to play the game optimally (Camerer, 2003; Rangel et al., 2008). Simple games therefore provide an excellent way to expose the computations underlying value-dependent choice in humans (Guth et al., 1982; Axelrod, 1984; Roth, 1995; Berg et al., 1995; Camerer, 2003; Glimcher et al., 2009; Montague et al., 2006).

In the domain of solution concepts, the idea of Nash equilibrium has been a leading principle (Fudenberg and Tirole, 1991; Nash, 1950). In a strategic interaction between two agents (two humans, a human and an institution, two institutions, etc.), the *Nash equilibria* are the set of choices by the two agents where no unilateral change by either agent can improve their outcome. More colloquially, in the absence of some other knowledge, if my choice is at a Nash equilibrium, then any other choice that I might have made does not improve my payoff. This statement also applies to my partner. The concept of Nash equilibrium is illustrated in Figure 1. In this case, we show the payoffs for each player in a table. For illustration purposes, we show the simple example of a simultaneous game where each player chooses their actions (there are two available to each player) independent of their partner, after which the choices and their attendant payoffs are revealed to both players.

Now consider the case where player 1 takes action 1 and player 2 takes action 2. This places each player in the upper right-hand cell of the payoff table. This is *not* a Nash equilibrium, since the outcome would improve for either player had they chosen the alternative action. This same conclusion would be true if player 1 had chosen action 2 and player 2 had chosen action 1, thus placing them in the lower left-hand cell of the table. In this illustration, the Nash equilibria are indicated by green circles. Let's suppose that the players make choices that place them in the upper left-hand cell. This set of choices is a Nash equilibrium because if either player were to unilaterally change their action, then their outcome gets worse. This is a very simple example, and we have made the actions mutually exclusive for illustration purposes (this is called a pure strategy Nash equilibrium). In some games, especially those that model pertinent real-world situations, a player can choose a mixed strategy where some fraction f of their choices are allocated to action 1 and the remaining fraction $1 - f$ of their choices are allocated to action 2. This is called a *mixed strategy* because a player's actions are a mixture of the available actions. The game shown in Figure 1 also possesses a mixed-strategy equilibrium; however, it

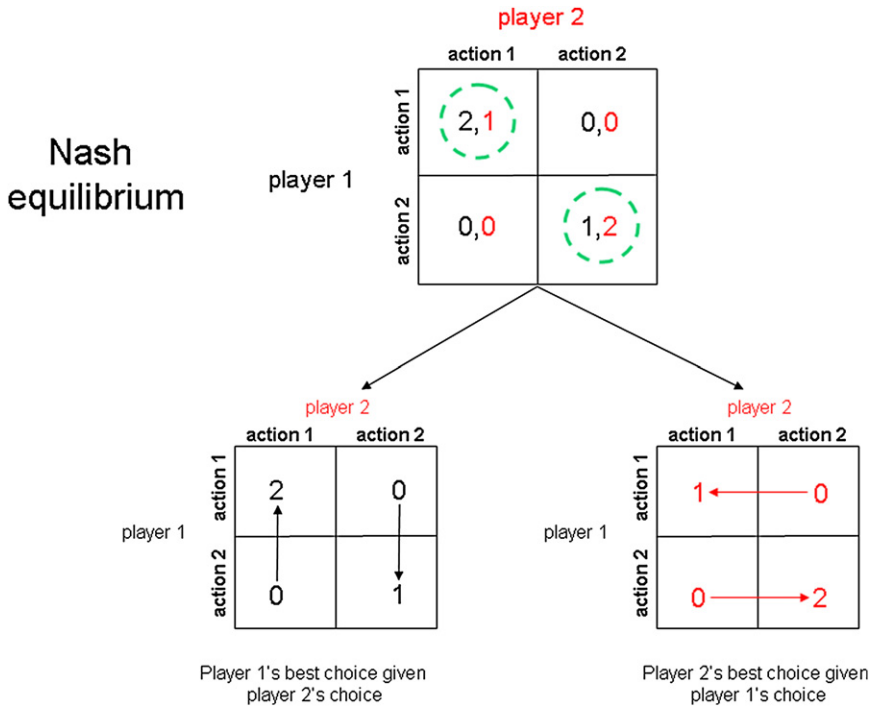

Figure 1. Nash Equilibrium

In a strategic interaction between two agents, the Nash equilibria are any of the set of choices made by the two agents where any unilateral change in strategy by one or the other agent does not improve the agent's outcome. Illustrated are payoff tables for a two-agent game where two agents (players) move simultaneously given two options (actions). (Top) The payoff table for any given set of choices made by the two agents. The green circles highlight the Nash equilibria. When the two players choose the same action (upper left or lower left quadrants) they are in a Nash equilibrium; in this state, any unilateral change in choice results in a worse outcome, and the choices are no longer in Nash Equilibrium. (Bottom) Payoff tables show the best choices for player 1 (left) given player 2's choices and for player 2 given player 1's choices. The set of players' choices in this example will tend to maintain Nash equilibrium (one of the two green circles), once it is discovered, since these provide the best possible outcomes for the pair.

exposes one flaw in this particular equilibrium concept: with the pure strategy one player always makes more than the other, and in the real world this might discourage participation altogether. The Nash concept shows what a player's best response should be provided that one's partner is playing their equilibrium strategy. In practice, partners do not always play their equilibrium strategy—humans often deviate from Nash in practice even when they know this to be suboptimal (van den Bos et al., 2008)—in these instances, playing according to a Nash equilibrium does not equate to the best response. These equilibrium concepts, and others, provide rational ways to identify the *best response* in a game, a central concept in economics (Fudenberg and Tirole, 1991).

Mixed-strategy games can be much better representations of real-world choice situations. For example, consider a weasel that visits a clearing possessing two holes. The weasel can poke its head into hole 1 and get water from an underground spring that flows sometimes to yield water or is otherwise dry. Or the weasel can poke its head into hole 2 where another animal has stashed nuts or not. How should the weasel allocate the fraction of its choices to each of these alternatives once it arrives in the clearing? In this case, each choice has variable yields, and the weasel's nervous system must decide the relative value of likely returns from each choice before nudging the weasel into a specific choice. Since the returns from these choices are stochastic, the weasel must bring a good model of the returns from each hole, and it must *continually re-estimate* those returns conditioned on its past experiences with each hole. This example shows that the weasel must compute a mixed strategy for the two choices, but it also illustrates the limit of equilibrium solutions to this game—the weasel must use some of its processing power to re-estimate the value of each choice and so simply focusing on equilibrium solutions would miss some of the central features of this real-world problem. Here, the weasel is constructing and updating predictive models of its (possibly dynamic) food sources, which represents a departure from simple equilibrium-based "best response" models.

This same reasoning would hold for humans interacting in real-world exchanges and suggests a kind of blending of strict equilibrium-based accounts of best responses and predictive models of others' behavior. Here, rational agents in the real world should estimate what fraction of the other agents will choose their equilibrium strategy and possess or guess something about what the nonequilibrium agents will do. These kinds of blended accounts (blends between economic models and predictive behavioral models) have given rise to cognitive hierarchy models of how human agents *should model* those with whom they interact (Camerer et al., 2004; also see Ray et al., 2008, for use of cognitive hierarchy model during social exchange).

Some of the most precise connections among valuation, choice, and neural function have been produced using game-theoretic settings in nonhuman primates paired with single-unit electrophysiological recordings (Dorris and Glimcher, 2004; Sugrue et al., 2004; Seo and Lee, 2007, 2009; also see Platt and Glimcher, 1999; Hayden et al., 2009; Hayden and Platt, 2010). Figure 2 shows three examples where the visual system was used primarily as an input-output device and the parameters of interest involved the valuation of "where to look." The experiment illustrated at the left represents work by Dorris and Glimcher using what is called the work-shirk inspection game. The animal gets one payoff if he is working when the "employer" looks in, another payoff if he is shirking his duties when the "employer" looks in, and so on. The best response for this game is a mixed strategy, that is, the animal should choose to distribute its actions across its two behavioral options

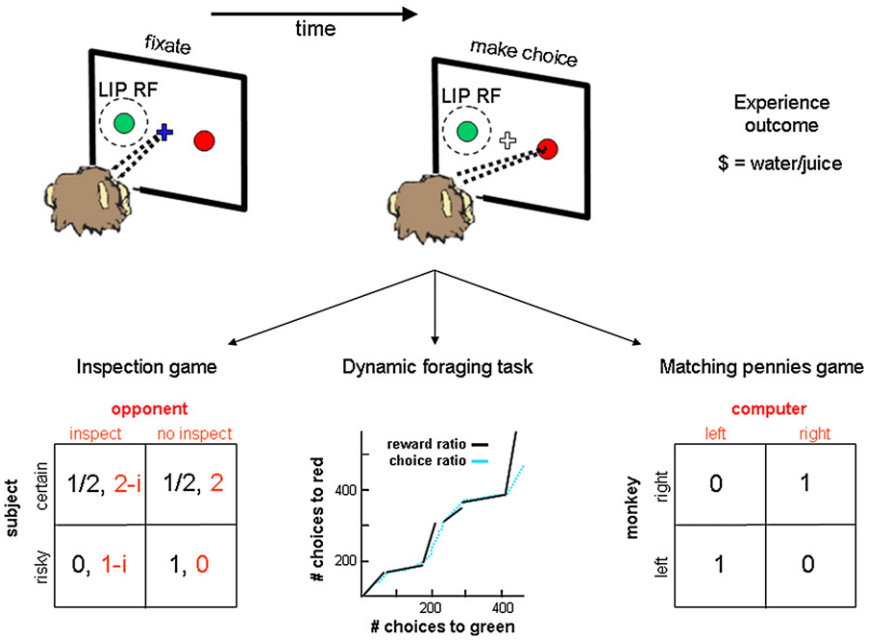

Figure 2. Neuroeconomic Approaches to Studying Choice Behavior in Nonhuman Primates Paired with Single-Unit Electrophysiological Recordings

(Top) Nonhuman primates are trained to perform choice tasks using eye movements: (left) monkeys fixate until cued to choose "action 1" or "action 2," which in this example is look to the red circle or look to the green circle, respectively. The circle placements are chosen based on the receptive field of the single unit under study, thus allowing study of the relationship between single-unit activity and the choice to take "action 1" or "action 2." In these studies, the reward is typical a squirt of juice or water, which is desirable for water-deprived (thirsty) players. (Bottom) Three game-theoretic settings used in nonhuman primates to investigate the connection between value, choice, and neural activity: (left) payoff table for a two-player *Inspection game*, (middle) *Dynamic foraging task*, and (right) *Matching pennies game* played between a monkey and a computer algorithm. These games and other uses of game-theoretic approaches are beginning to expose and model expected changes in subjects' behavior.

(e.g., 20% to one option, 80% to the other) depending on the setting of a parameter. This game took place while the experimenters recorded from neurons in the posterior parietal cortex. The clever maneuver in this experiment is that the experimenters parameterized the Nash equilibrium in this game (controlled by parameter i in Figure 2, left panel). The main conclusion in that study was that subjective desirability of a behavioral option covaried with firing-rate changes in the recorded neurons independent of the objective parameters related to reward acquisition. This explanation of the observed data has been questioned (Sugrue et al., 2005) by an alternative experiment that manipulated local changes in a reward-harvesting task (Sugrue et al., 2004, see below) while monitoring neural activity in the same region. We highlight these examples (Dorris and Glimcher, 2004; Sugrue et al., 2004) to demonstrate that the nature of the quantitative economic choice model in constructing an experiment is very important. Similar approaches will continue to improve our model-based understanding of the decision variables encoded in recordable neural activity.

A similar kind of game-theory strategy was used by Lee and colleagues where a "matching pennies" game (a coordination game like the example in Figure 2, right panel) was used in monkeys while single-unit recordings were carried out in the brain (Barraclough et al., 2004; Seo and Lee, 2007). Again, correlations between neural activity and variables that in theory could (or should) influence outcomes were observed. Sugrue et al. (2004) demonstrated another experiment that has used a game-like probe while recording neural activity. This group used a visual reward-harvesting task and also recorded from neurons in the parietal cortex while the rate of reward from different behavioral options was controlled (a dynamic foraging task akin to the example in Figure 2, middle panel). These investigators found that their recordings were most consistent with the relative value of competing options (here local probability of eye movements to one of two targets) rather than subjective desirability (the average payoff of each target). These examples represent a small fraction of an ever-growing literature using game-theory-designed tasks to probe animals while recording some kind of neural variable.

This work is in its early days; however, one common theme emerges from these experiments, that is, the ability to use game-theoretic probes to expose and model expected changes in subjects' *behavior*. While the connection of economic variables to single-neuron activity in these studies remains either provisional or in some cases disputed, there is no dispute that the experimental probes provide an excellent way to probe value-dependent choice in primates at the *behavioral level*. This conclusion is supported by the fact that all three studies highlighted in Figure 2 produced excellent and quite sensitive behavioral models of the animal's observed choice behavior (Dorris and Glimcher, 2004; Sugrue et al., 2004; Seo and Lee, 2007). The conflicting results (Dorris and Glimcher, 2004; Sugrue et al., 2004) may simply be a limitation of the current experimental capacity to record from a sufficient number and range of neurons during tasks. Alternatively, it may reflect a deeper issue of the myriad ways that different brains implement solutions to economic problems at the level of neural networks. Nevertheless, the behavioral lesson in the animal work has been taken to heart in the human neurobiology community where detailed game-theoretic probes have been used to probe everything from response to monetary reward, risk, and even response to the risks involved in exchanging with other humans.

Interpersonal Economic Exchange and Its Applications

Social exchange occurs in species ranging from insects to humans (Hamilton, 1964a, 1964b; Trivers, 1971). In primates,

reciprocal social interactions with non-kin occur repeatedly, thus necessitating the capacity to assign social credit or blame for shared outcomes and to act appropriately according to these assignments (Maynard Smith and Price, 1973; Axelrod and Hamilton, 1981; Nowak and Sigmund, 1992). In humans, reciprocity is one of a collection of mechanisms necessary to support social exchange (e.g., Trivers, 1971); yet, the underlying neural representations of these mechanisms remain murky. In almost all social exchanges, one must detect and accurately track which social agent (who) gets credit for an outcome. Should credit for an outcome be assigned to one's own actions or those of one's partner? Understanding such agent-specific computations is important, because the assignment of social agency breaks down in a range of mental illnesses, including schizophrenia and autism spectrum disorders (*assignment of social agency*: Frith and Frith, 2001; Vogeley et al., 2001; Kelley et al., 2002; Lieberman and Pfeifer, 2005; Ochsner et al., 2004; Seger et al., 2004; Vogeley and Fink, 2003; Decety and Sommerville, 2003; Lieberman, 2007; Mitchell et al., 2006; Northoff et al., 2006; Ochsner et al., 2005; Saxe et al., 2004; Uddin et al., 2007; *breakdown*: Baron-Cohen and Belmonte, 2005; Frith and Frith, 1999; Brüne, 2005; Wischniewski et al., 2009). Social agency computations are also a prerequisite for generating models of others' mental states. This latter capacity, called theory of mind, is highly developed in humans and has been shown to activate a consistent set of brain regions in neuroimaging experiments (Gallagher et al., 2000; Brunet et al., 2000; Wicker et al., 2003; Decety et al., 2004). But establishing agency, while an important underlying computation for social exchange, is not sufficient to interact fruitfully with others.

Social exchange, even in its simplest settings, requires a collection of important computational capacities in the minds of the interacting agents. In a simple *fair* trade with another human, the brain of each trader must be able to (1) *compute* norms for what is considered fair, (2) *detect* deviations from such norms, and (3) *possess* the capacity to select appropriate actions based on these deviations. An adaptive social creature must understand the meaning of social gestures and the "normative" responses expected from those gestures—all the while trying to model the specific agent with whom they are currently interacting. This is a ferociously difficult problem, and some of the most important pathologies of our mental life revolve around perturbed function in the domain of social interaction, including schizophrenia, depression, social anxiety, autism, and personality disorders. Therefore, it is fundamentally important to try and develop objective, quantitative measures of the neural and behavioral underpinnings of social exchange.

Recent work has begun to identify computational models of social agent detection, social signaling, and social learning that connect to measurable neural responses in humans (for review, see Behrens et al., 2009). This nascent area represents an opportunity for building computational descriptions of social exchange that can connect to underlying neurobiology, in ways that inform both the behavioral and neural computations underlying pathologies of social interaction seen in a variety of psychiatric illnesses. Mathematically portrayed economic exchange games have dominated this new and quantitative approach to social exchange (Delgado et al., 2005; King-Casas et al., 2005; Rilling et al., 2002, 2004; Sanfey et al., 2003; Singer et al., 2006) because they assess a subject's internal norm for what is fair in an exchange, and they require that each subject model their partner's mental state (Camerer, 2003; Camerer and Fehr, 2006; Kagel and Roth, 1997; Montague and Lohrenz, 2007). In summary, these games expose the three component computations required above for normal social exchange between two agents. The games have a variety of names—the well-known prisoner's dilemma, the dictator game, the ultimatum game, the trust game, and so on (Axelrod, 1984; Guth et al., 1982; Roth, 1995; Camerer, 2003). They are excellent experimental probes because they are simple and mathematically well-specified, there is an existing body of behavioral data employing them across a variety of contexts, and there are known solution concepts for how they "should" be played by a rational self-interested agent (Roth, 1995; Camerer, 2003; Camerer and Fehr, 2006). More importantly, they all require participants to model their partner. Such simplified behavioral probes provide an excellent starting point for extracting quantitative descriptions of social signaling and its pathologies because the parameter space is manageable, and reasonable normative solutions to these games exist (Camerer, 2003; but see Greenwald and Jafari, 2003, for complications in solution concepts). These games have already proven valuable in early work in clinical populations. The power of these games resides in their payoff structures and the fact that they could be used to decompose social interactions into a collection of behavioral and neural primitives. Such primitives could provide a new lexicon useful for assessing medical or behavior therapies, understanding differing neural strategies for decision-making, and so on.

Single-Shot Fairness Games

The ultimatum game involves two players—the proposer and the responder (Guth et al., 1982)—and could reasonably be renamed "take-it-or-leave-it." In this game, the proposer is endowed with some resource (say $100) and can offer any split to the responder. Let's suppose the proposer offers $80 for herself and $20 for the responder. If the responder accepts the split, then both players walk away with money ("take it" option). If the responder rejects, neither player gets anything ("leave it" option). Rationally, the proposer should send $1, and the responder should accept all non-zero offers since they start with nothing, but experiments show this expectation to be false. In practice, the proposer sends $40 as their modal offer and responders reject 50% of the time at an $80:$20 split. These behavioral results are remarkably consistent and do not appreciably change across diverse cultural settings and experimental arrangements (Heinrich, 2004). This simple one-shot game provides an excellent tool to probe the detection of and response to fairness deviations in an all-or-none fashion; however, one-shot games miss one central feature of all social exchange—learning. The single-shot games do not provide an opportunity to see the effects of the social signal sent to one's partner. Humans routinely send social signals to one another (encoded in a variety of currencies) with the expectation of adjusting their partner's behavior in future interactions. Importantly, cooperation in repeated interactions (relationships) can wax and wane, and social signaling and learning enables individuals to fix

relationships when they falter. Thus, the limitations posed by single-shot interactions can be overcome in part by using a multi-round fairness game.

Interpersonal Fairness Games and the Computations They Expose

Fairness games collectively expose three important classes of computations that must be operable during successful two-party social exchanges:

1) Humans Compute or Retrieve Stable Shared Norms for What Is Expected in Reciprocal Trades between Two Individuals

The large initial offers in the ultimatum game, the stable rejection rate (50%) at an 80:20 split, and the lack of dependence on cultural factors (Heinrich, 2004) strongly support the existence of fairness norms for offers during a two-party exchange. Such norms provide baseline (prior) distributions of acceptable signals to be sent to others and whether signals received from others are acceptable. It has been suggested recently that such norms may also be part of a computational depiction of emotion processing (Montague and Lohrenz, 2007), but this area to date remains undeveloped. In any case, the presence of norms and their use in decision-making is parametrically revealed by fairness games.

2) Humans Possess Sensitive Norm-Violation Detection Mechanisms and These Violations Represent Natural Classes of Learning Signals

To respond to norm violations, humans must generate and respond to error signals carrying information about the norm violation. These kinds of signals have been hinted at in the reward-processing literature, especially as it makes contact with dopaminergic systems. A large subset of midbrain dopamine neurons participate in circuits that learn to value and to predict future rewarding events, especially the delivery of primary rewards like food, water, and sex (Montague et al., 1996; Schultz et al., 1997; Schultz, 1998; Schultz and Dickinson, 2000; Dayan and Abbott, 2001; Montague and Berns, 2002; Bayer and Glimcher, 2005; Montague et al., 2004, 2006). More specifically, midbrain dopamine neurons are thought to emit reward-prediction errors encoded in modulations in their spike output (Montague et al., 1996; for review see Schultz, 1998; Schultz and Dickinson, 2000). This interpretation is strongly supported for the timing of burst and pause responses in the spike trains of these neurons (Schultz, 1998; Bayer and Glimcher, 2005). In this work, the goal is to learn from the statistical structure of reward delivery actually experienced to determine when expectations have been violated (see Montague et al., 2004, for review). However, for a "fair" exchange with another human, a control signal for norm violation would need to be able to report to a creature that a social partner's behavior differed from typical behavior of social partners, both in the mean (like the reward prediction error) and the variance (like an error in uncertainty signal), and possibly even include errors in higher-order moments (see Montague and Lohrenz, 2007).

3) Humans Are Willing, at a Cost to Themselves, to Send a Corrective Signal (Learning Signal) to Their Partner

To understand the kind of social signal to send to a partner with the goal of engineering their behavior, a subject must have a sufficient model of the partner's likely response to the signal. In addition, the willingness to send such a signal at a cost shows that humans know the likely "return" on the signal. The fairness games as outlined above and described in detail below expose this class of computation. So this third capacity requires a good estimate of the partner's response and the likely return—although economically framed in this description, these are deep social signaling and estimation capacities. Also, in an experiment by Sanfey et al. (2003), a human's propensity to reject unfair splits and the attendant brain responses to such unfair offers depend on whether they are playing a human or a computer. Consequently, computation three depends on the nature of the partner, which is reasonable since corrective signals sent in such exchanges only work in a particular kind of agent—a human.

In summary, the use of fairness games provides a window onto these three classes of computation that take place in simple two-party exchanges. And while such probes are necessarily simplified, they are quantitative and general enough to provide insights into some of the most important component parts of social exchange.

Biomarkers Exposed by Fairness Games

Fairness games are increasingly being used to investigate pathological social behavior associated with a variety of psychiatric illnesses, including borderline personality disorder (King-Casas et al., 2008; Seres et al., 2009; Unoka et al., 2009), psychopathy (Mokros et al., 2008; Rilling et al., 2007), social anxiety (Sripada et al., 2009), depression (Hokanson et al., 1980), addiction (Yi et al., 2007), and autism (Chiu et al., 2008a; Andari et al., 2010; Yoshida et al., 2010). While each of these illnesses confers significant social impairment, neurobiological research has traditionally overlooked social symptoms of psychiatric illness, in part due to an assumption that behavioral and neural computations underlying social behaviors are too complex to be amenable to rigorous study. However, the recent work using fairness games suggests the promise for these probes of social behavior to be fruitful in providing sensitive and specific biomarkers for social pathologies.

One example of this approach has used an iterated "trust game" to investigate social symptoms of borderline personality disorder (Figure 3; King-Casas et al., 2008; Seres et al., 2009; Unoka et al., 2009). Similar to the ultimatum game discussed above, the trust game relies on a shared sense of fairness in order for an exchange to be mutually beneficial (Berg et al., 1995; Camerer and Weigelt, 1988). In this game, one person, dubbed the "investor," is endowed with a valued resource, typically money (Figure 3A). An investor can then send any portion of that endowment (or nothing at all) to a social partner. Whatever portion of the endowment the investor chooses to send is automatically tripled before being received by the "trustee," who then has the opportunity to repay some portion of the tripled investment. Critically, both the size of investment and size of repayment are entirely at the discretion the sender; that is, neither investor nor trustee is required to send anything. Thus, *trust* can be quantified as the amount of money one person sends to another without external enforcement. If both players

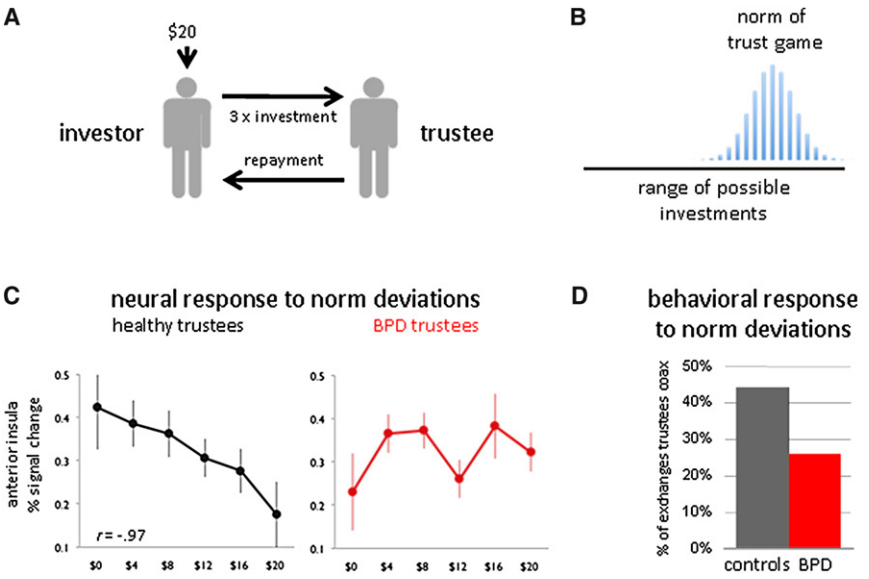

Figure 3. Biomarker for Borderline Personality Disorder Exposed by Multiround Trust Game

(A) Iterated trust game: investors are endowed with $20 at the start of each of ten rounds. Investors can invest any portion of the endowment with their partner. The invested amount is tripled before being passed to the trustee, who can repay any portion of the tripled investment.

(B) The social norm for investor behavior in the trust game can be represented as distributions of likely investments given prior expectations and accrued experience. Investments are typically large in this game of cooperation. Thus, small investments represent a deviation from the social norm of the task.

(C) A region-of-interest analysis of anterior insula of 38 healthy trustees shows hemodynamic activity to be negatively related to size of investment, consistent with the idea that small investments represent a deviation in norm. In contrast, similar analyses among 55 individuals with BPD showed no significant activation in these regions when investments were small, suggesting that low investments do not represent a violation of an established norm.

(D) Healthy trustees are twice as likely as BPD trustees to coax when cooperation between players is low. Specifically, healthy trustees are more likely to make a large repayment ($\geq$ investment amount) after having received a small investment. Conversely, BPD trustees are more likely to make a small repayment (< investment amount) after receiving a small investment.

in this game share and act upon a common social norm, for example that they ought to share the winnings of a game equally (Fehr and Schmidt, 1999), then an optimal shared strategy would be for the investor to send the entire endowment (e.g., invest $20 of $20) and for the trustee to send back half of the tripled investment (repay $30 of $60). In the context of this structured exchange, a shared norm and cooperative strategy mutually benefit both players.

In King-Casas et al. (2008), we used a multiround version of this simple trust exchange to examine the trajectory of cooperation between healthy investors and trustees with borderline personality disorder. Individuals with borderline personality disorder typically have difficulty navigating social relationships, exhibiting a pattern of unstable and intense interpersonal relationships, frantic efforts to avoid abandonment, inappropriate, intense anger or difficulty controlling anger, and affective instability (American Psychiatric Association, 2000). Thus, we sought biomarkers of aberrant social norms and aberrant responses to norm violations in the context of the repeated trust game, in order to distinguish healthy cooperative exchange in control subjects from abnormal social interaction among individuals with borderline personality disorder.

Although initial levels of cooperation were found to be comparable between healthy dyads (n = 38; healthy investors and healthy trustee) and BPD dyads (n = 55; healthy investors playing BPD trustee), cooperation in BPD dyads broke down across the repeated exchange. Importantly, the breakdown in cooperation was not attributable to diagnostic differences in the use of reciprocity as a behavioral strategy (see King-Casas et al., 2005), indicating that individual BPDs are sensitive to and respond appropriately to the behavioral signals emitted by their partners. Rather, diagnostic differences were found when cooperation between players began to falter. In particular, when investors sent small investments to healthy trustees, healthy trustees responded neurally with greater hemodynamic activity in anterior insula and responded behaviorally by sending unusually large repayments (Figure 3C). This "coaxing" strategy was effective in signaling their own trustworthiness and typically led to greater investments on subsequent rounds. In contrast, individuals with BPD showed no increased activity in anterior insula and were half as likely to "coax" back higher investments through generous repayments (Figure 3D).

These results can be interpreted through recent work in affective, interoceptive, economic, and social domains that implicate anterior insular cortex in the representation and updating of norms (Montague and Lohrenz, 2007). For example, in the ultimatum game discussed above, anterior insula activity both scales negatively with offer size (greater activity to unfair offers) and predicts whether the offer is subsequently rejected (Sanfey et al., 2003). Similarly, anterior insula activity of observers is greater when a punishment is applied to players perceived as fair relative to players perceived as unfair (Singer et al., 2006). While in nonsocial decision-making tasks, activity in the anterior insula both encodes and updates representations of risk about decision outcomes (Preuschoff et al., 2006, 2008).

Taken together, insula responses related to low investment levels in healthy control trustees likely reflect these subjects' sensitivity to deviations in normative behavior of a trust game. That is, healthy control trustees expect greater investment levels from social partners, and seek to "coax" back high investments when cooperation begins to break down. In contrast, trustees with borderline personality disorder show decreased sensitivity to the social norm of the game (high investments) and were therefore less likely to coax back the normative behavior from their partner. This result suggests that borderline personality disorder confers a diminished capacity to represent

expectations for social partners, and as a consequence individuals with BPD cannot take corrective action (social control signal) that might serve to reestablish cooperative interaction.

While the diagnostic and predictive utility of such approaches to understanding social pathologies remains to be proved, the sensitivity of behavior in fairness games to modulation by common pharmacological interventions suggests that identified aberrant social computations may be effectively treated with already available (and developing) therapeutics (e.g., *serotonergic modulation*: Crockett et al., 2008; Tse and Bond, 2002a, 2002b; Wood et al., 2006; *oxytocin modulation*: Kosfeld et al., 2005; Baumgartner et al., 2008; Declerck et al., 2010; Andari et al., 2010).

Is There a Genetic Basis for the Computations Carried Out in Fairness Games?

The biomarkers exposed by interpersonal exchange games could provide quantitative phenotypes that geneticists could use to identify genes responsible for building neural circuitry with the capacity to make neurotypical computations and therefore express neurotypical decision behavior. Furthermore, genes underlying subtle to extreme disruptions in the capacity to perform these computations could be identified and would serve as the ultimate biomarkers of mental disorders. The most common mental disorders have been demonstrated to be highly heritable, including schizophrenia, bipolar disorder, major depression, and autism spectrum disorder; however, only recently has evidence emerged that suggests traits expressed in fairness games and neuroimaging experiments are heritable as well. Recent work has demonstrated high heritability of fairness game behavior (Cesarini et al., 2008; Wallace et al., 2007) and fMRI responses (Matthews et al., 2007) in humans. Specifically, cooperative behavior (investment and reciprocity measures) in the single-shot trust game was demonstrated to be highly influenced by genetic factors (Cesarini et al., 2008) as were rejection rates in the ultimatum game (Wallace et al., 2007). Additionally, polymorphisms in the oxytocin receptor gene (OXTR) have been associated with specific behavioral responses in a dictator game (Israel et al., 2009).

The heritability reported in this work matches or is better than estimated heritability using DSM IV criteria; however, economic games are quantitative and parametric while the DSM uses categorical criteria often related directly to the symptom lists that define the disorders. Consequently, the economic games may provide a new way to better understand the successes of DSM IV criteria, identify its failures, and provide new ways to look for genetic underpinnings. Finally, brain responses measured with fMRI during a conflict response task were recently shown to be heritable as well in a study involving female twin pairs (Matthews et al., 2007). Together, these studies suggest that fairness games and the computations they expose can generate quantitative phenotypes that will allow for further investigation into the underlying genetic influences of decision-making in neurotypical individuals and individuals with mental disorders.

Summary

Social exchange is a gamble all humans take; however, humans mitigate the risks associated with such exchanges by relying on deep models of other humans' likely intentions and responses. While lifelong experience with others humans clearly influences these models of others, it is also likely that many important components of such models are inherited from our ancestors in order to provide good priors for estimating the risks and rewards for engaging in a wide range of interpersonal exchanges. When the biological substrates implementing these models are damaged or altered in a significant way, abnormal behavior is certain to be expressed. Economic games are beginning to provide new ways to capture and quantify this behavior and its associated neural correlates and may well produce new biomarkers of mental disease. Furthermore, these probes seem ideally suited to identify some of the genetic underpinnings of important quantitative features of mental function because they generate parametric variation along a variety of cognitive dimensions. This last point is quite important since many aspects of mental illness express as quantitative differences along normal cognitive dimensions. This work is still in its infancy, and so the real payoffs lie mainly in the future. However, the integration of economically framed probes with modern measures of neural function appears to be a growth area especially in the possible applications to mental disease and brain injury.

REFERENCES

American Psychiatric Association. (2000). Diagnostic and Statistical Manual of Mental Disorders: DSM-IV (Washington, DC: American Psychiatric Association).

Andari, E., Duhamel, J.R., Zalla, T., Herbrecht, E., Leboyer, M., and Sirigu, A. (2010). Promoting social behavior with oxytocin in high-functioning autism spectrum disorders. Proc. Natl. Acad. Sci. USA *107*, 4389–4394.

Axelrod, R. (1984). The Evolution of Cooperation (New York: Basic Books).

Axelrod, R., and Hamilton, W.D. (1981). The evolution of cooperation. Science *211*, 1390–1396.

Baron-Cohen, S., and Belmonte, M.K. (2005). Autism: a window onto the development of the social and the analytic brain. Annu. Rev. Neurosci. *28*, 109–126.

Barraclough, D.J., Conroy, M.L., and Lee, D. (2004). Prefrontal cortex and decision making in a mixed-strategy game. Nat. Neurosci. *7*, 404–410.

Baumgartner, T., Heinrichs, M., Vonlanthen, A., Fischbacher, U., and Fehr, E. (2008). Oxytocin shapes the neural circuitry of trust and trust adaptation in humans. Neuron *58*, 639–650.

Bayer, H.M., and Glimcher, P.W. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. Neuron *47*, 129–141.

Behrens, T.E., Hunt, L.T., and Rushworth, M.F. (2009). The computation of social behavior. Science *324*, 1160–1164.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. Games Econ. Behav. *10*, 122–142.

Bickel, W.K., Yi, R., Kowal, B.P., and Gatchalian, K.M. (2008). Cigarette smokers discount past and future rewards symmetrically and more than controls: is discounting a measure of impulsivity? Drug Alcohol Depend. *96*, 256–262.

Brüne, M. (2005). "Theory of mind" in schizophrenia: a review of the literature. Schizophr. Bull. *31*, 21–42.

Brunet, E., Sarfati, Y., Hardy-Bayle, M.C., and Decety, J. (2000). A PET investigation of the attribution of intentions with a nonverbal task. Neuroimage *11*, 157–166.

Camerer, C.F. (2003). Behavioral Game Theory: Experiments in Strategic Interaction (Princeton, NJ: Princeton University Press).

Camerer, C.F., and Fehr, E. (2006). When does "economic man" dominate social behavior? Science *311*, 47–52.

Camerer, C.F., and Ho, T.H. (1999). Experience-weighted attraction learning in normal form games. Econometrica *67*, 827–874.

Camerer, C., and Weigelt, K. (1988). Experimental tests of a sequential equilibrium model. Econometrica *56*, 1–36.

Camerer, C.F., Ho, T.H., and Chong, J.K. (2004). A cognitive hierarchy model of games. Q. J. Econ. *119*, 861–898.

Cesarini, D., Dawes, C.T., Fowler, J.H., Johannesson, M., Lichtenstein, P., and Wallace, B. (2008). Heritability of cooperative behavior in the trust game. Proc. Natl. Acad. Sci. USA *105*, 3721–3726.

Chiu, P.H., Kayali, M.A., Kishida, K.T., Tomlin, D., Klinger, L.G., Klinger, M.R., and Montague, P.R. (2008a). Self responses along cingulate cortex reveal quantitative neural phenotype for high-functioning autism. Neuron *57*, 463–473.

Chiu, P.H., Lohrenz, T.M., and Montague, P.R. (2008b). Smokers' brains compute, but ignore, a fictive error signal in a sequential investment task. Nat. Neurosci. *11*, 514–520.

Clark, L., Bechara, A., Damasio, H., Aitken, M.R., Sahakian, B.J., and Robbins, T.W. (2008). Differential effects of insular and ventromedial prefrontal cortex lesions on risky decision-making. Brain *131*, 1311–1322.

Critchley, H.D., Mathias, C.J., and Dolan, R.J. (2001). Neural activity in the human brain relating to uncertainty and arousal during anticipation. Neuron *29*, 537–545.

Crockett, M.J., Clark, L., Tabibnia, G., Lieberman, M.D., and Robbins, T.W. (2008). Serotonin modulates behavioral reactions to unfairness. Science *320*, 1739.

Dayan, P., and Abbott, L.F. (2001). Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems (Cambridge, MA: MIT Press).

Decety, J., and Sommerville, J.A. (2003). Shared representations between self and other: a social cognitive neuroscience view. Trends Cogn. Sci. *7*, 527–533.

Decety, J., Jackson, P.L., Sommerville, J.A., Chaminade, T., and Meltzoff, A.N. (2004). The neural bases of cooperation and competition: an fMRI investigation. Neuroimage *23*, 744–751.

Declerck, C.H., Boone, C., and Kiyonari, T. (2010). Oxytocin and cooperation under conditions of uncertainty: the modulating role of incentives and social information. Horm. Behav. *57*, 368–374.

Delgado, M.R., Frank, R.H., and Phelps, E.A. (2005). Perceptions of moral character modulate the neural systems of reward during the trust game. Nat. Neurosci. *8*, 1611–1618.

Delgado, M.R., Nearing, K.I., Ledoux, J.E., and Phelps, E.A. (2008). Neural circuitry underlying the regulation of conditioned fear and its relation to extinction. Neuron *59*, 829–838.

DeMartino, B., Camerer, C.F., and Adolphs, R. (2010). Amygdala damage eliminates monetary loss aversion. Proc. Natl. Acad. Sci. USA *107*, 3788–3792.

Dorris, M.C., and Glimcher, P.W. (2004). Activity in posterior parietal cortex is correlated with the relative subjective desirability of action. Neuron *44*, 365–378.

Fehr, E., and Schmidt, K.M. (1999). A theory of fairness, competition, and cooperation. Q. J. Econ. *114*, 817–868.

Feinstein, J.S., Stein, M.B., and Paulus, M.P. (2006). Anterior insula reactivity during certain decisions is associated with neuroticism. Soc. Cogn. Affect. Neurosci. *1*, 136–142.

Frith, C.D., and Frith, U. (1999). Interacting minds—a biological basis. Science *286*, 1692–1695.

Frith, U., and Frith, C. (2001). The biological basis of social interaction. Curr. Dir. Psychol. Sci. *10*, 151–155.

Fudenberg, D., and Tirole, J. (1991). Game Theory (Cambridge, MA: MIT Press).

Gallagher, H.L., Happe, F., Brunswick, N., Fletcher, P.C., Frith, U., and Frith, C.D. (2000). Reading the mind in cartoons and stories: an fMRI study of 'theory of mind' in verbal and nonverbal tasks. Neuropsychologia *38*, 11–21.

Glimcher, P.W., and Rustichini, A. (2004). Neuroeconomics: the consilience of brain and decision. Science *306*, 447–452.

Glimcher, P.W., Camerer, C., Poldrack, R.A., and Fehr, E. (2009). Neuroeconomics: Decision Making and the Brain (New York, NY: Academic Press).

Greenwald, A., and Jafari, A. (2003) A class of no-regret algorithms and game-theoretic equilibria. In COLT '03: Proceedings of the 16th Conference on Computational Learning Theory, pp 1–11.

Guth, W., Schmittberger, R., and Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. J. Econ. Behav. Organ. *3*, 367–388.

Hamilton, W.D. (1964a). The genetical evolution of social behaviour. I. J. Theor. Biol. *7*, 1–16.

Hamilton, W.D. (1964b). The genetical evolution of social behaviour. II. J. Theor. Biol. *7*, 17–52.

Hayden, B.Y., and Platt, M.L. (2010). Neurons in anterior cingulate cortex multiplex information about reward and action. J. Neurosci. *30*, 3339–3346.

Hayden, B.Y., Pearson, J.M., and Platt, M.L. (2009). Fictive reward signals in the anterior cingulate cortex. Science *324*, 948–950.

Heinrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. J. Econ. Behav. Organ. *53*, 3–35.

Hokanson, J.E., Sacco, W.P., Blumberg, S.R., and Landrum, G.C. (1980). Interpersonal behavior of depressive individuals in a mixed-motive game. J. Abnorm. Psychol. *89*, 320–332.

Hsu, M., Bhatt, M., Adolphs, R., Tranel, D., and Camerer, C.F. (2005). Neural systems responding to degrees of uncertainty in human decision-making. Science *310*, 1680–1683.

Huettel, S.A., Stowe, C.J., Gordon, E.M., Warner, B.T., and Platt, M.L. (2006). Neural signatures of economic preferences for risk and ambiguity. Neuron *49*, 765–775.

Israel, S., Lerer, E., Shalev, I., Uzefovsky, F., Riebold, M., Laiba, E., Bachner-Melman, R., Maril, A., Bornstein, G., Knafo, A., and Ebstein, R.P. (2009). The oxytocin receptor (OXTR) contributes to prosocial fund allocations in the dictator game and the social value orientations task. PLoS ONE *4*, e5535.

Kagel, J.H., and Roth, A.E. (1997). The Handbook of Experimental Economics (Princeton, NJ: Princeton University Press).

Kahn, H. (1960). On Thermonuclear War (Princeton, NJ: Princeton University Press).

Kahneman, D., and Tversky, A. (2000). Choices, Values and Frames (New York: Cambridge University Press and the Russell Sage Foundation)., 673–692.

Kelley, W.M., Macrae, C.N., Wyland, C.L., Caglar, S., Inati, S., and Heatherton, T.F. (2002). Finding the self? An event-related fMRI study. J. Cogn. Neurosci. *14*, 785–794.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C.F., Quartz, S.R., and Montague, P.R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. Science *308*, 78–83.

King-Casas, B., Sharp, C., Lomax-Bream, L., Lohrenz, T., Fonagy, P., and Montague, P.R. (2008). The rupture and repair of cooperation in borderline personality disorder. Science *321*, 806–810.

Kosfeld, M., Heinrichs, M., Zak, P.J., Fischbacher, U., and Fehr, E. (2005). Oxytocin increases trust in humans. Nature *435*, 673–676.

Lee, D. (2005). Neuroeconomics: making risky choices in the brain. Nat. Neurosci. *8*, 1129–1130.

Leland, D.S., and Paulus, M.P. (2005). Increased risk-taking decision-making but not altered response to punishment in stimulant-using young adults. Drug Alcohol Depend. *78*, 83–90.

Leland, D.S., Arce, E., Feinstein, J.S., and Paulus, M.P. (2006). Young adult stimulant users' increased striatal activation during uncertainty is related to impulsivity. Neuroimage *33*, 725–731.

Lieberman, M.D. (2007). Social cognitive neuroscience: a review of core processes. Annu. Rev. Psychol. *58*, 259–289.

Lieberman, M.D., and Pfeifer, J.H. (2005). The self and social perception: three kinds of questions in social cognitive neuroscience. In Cognitive Neuroscience of Emotional and Social Behavior, A. Easton and N. Emery, eds. (Philadephia, PA: Psychol. Press), pp. 195–235.

Lohrenz, T., McCabe, K., Camerer, C.F., and Montague, P.R. (2007). Neural signature of fictive learning signals in a sequential investment task. Proc. Natl. Acad. Sci. USA *104*, 9493–9498.

Ludewig, S., Paulus, M.P., Ludewig, K., and Vollenweider, F.X. (2003). Decision-making strategies by panic disorder subjects are more sensitive to errors. J. Affect. Disord. *76*, 183–189.

Matthews, S.C., Simmons, A.N., Strigo, I., Jang, K., Stein, M.B., and Paulus, M.P. (2007). Heritability of anterior cingulate response to conflict: an fMRI study in female twins. Neuroimage *38*, 223–227.

Maynard-Smith, J. (1982). Evolution and the Theory of Games (Cambridge, UK: Cambridge University Press).

Maynard Smith, J., and Price, G.R. (1973). The logic of animal conflict. Nature *246*, 15–18.

Minassian, A., Paulus, M.P., and Perry, W. (2004). Increased sensitivity to error during decision-making in bipolar disorder patients with acute mania. J. Affect. Disord. *82*, 203–208.

Minassian, A., Paulus, M., Lincoln, A., and Perry, W. (2007). Adults with autism show increased sensitivity to outcomes at low error rates during decision-making. J. Autism Dev. Disord. *37*, 1279–1288.

Mitchell, J.P., Macrae, C.N., and Banaji, M.R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. Neuron *50*, 655–663.

Mokros, A., Menner, B., Eisenbarth, H., Alpers, G.W., Lange, K.W., and Osterheider, M. (2008). Diminished cooperativeness of psychopaths in a prisoner's dilemma game yields higher rewards. J. Abnorm. Psychol. *117*, 406–413.

Montague, P.R., and Assad, J. (2008). Editorial overview. Curr. Opin. Neurobiol. *18*, 117–119.

Montague, P.R., and Berns, G.S. (2002). Neural economics and the biological substrates of valuation. Neuron *36*, 265–284.

Montague, P.R., and Lohrenz, T. (2007). To detect and correct: norm violations and their enforcement. Neuron *56*, 14–18.

Montague, P.R., Dayan, P., and Sejnowski, T.J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. J. Neurosci. *16*, 1936–1947.

Montague, P.R., Hyman, S.E., and Cohen, J.D. (2004). Computational roles for dopamine in behavioural control. Nature *431*, 760–767.

Montague, P.R., King-Casas, B., and Cohen, J.D. (2006). Imaging valuation models in human choice. Annu. Rev. Neurosci. *29*, 417–448.

Nash, J.F. (1950). Equilibrium points in n-person games. Proc. Natl. Acad. Sci. USA *36*, 48–49.

Niederle, M., and Roth, A.E. (2003). Relationship between wages and presence of a match in medical fellowships. JAMA *290*, 1153–1154.

Northoff, G., Heinzel, A., de Greck, M., Bermpohl, F., Dobrowolny, H., and Panksepp, J. (2006). Self-referential processing in our brain–a meta-analysis of imaging studies on the self. Neuroimage *31*, 440–457.

Nowak, M., and Sigmund, K. (1992). Tit for tat in heterogeneous populations. Nature *355*, 250–253.

Ochsner, K.N., Knierim, K., Ludlow, D.H., Hanelin, J., Ramachandran, T., Glover, G., and Mackey, S.C. (2004). Reflecting upon feelings: an fMRI study of neural systems supporting the attribution of emotion to self and other. J. Cogn. Neurosci. *16*, 1746–1772.

Ochsner, K.N., Beer, J.S., Robertson, E.R., Cooper, J.C., Gabrieli, J.D., Kihsltrom, J.F., and D'Esposito, M. (2005). The neural correlates of direct and reflected self-knowledge. Neuroimage *28*, 797–814.

Paulus, M.P., Geyer, M.A., and Braff, D.L. (1996). Use of methods from chaos theory to quantify a fundamental dysfunction in the behavioral organization of schizophrenic patients. Am. J. Psychiatry *153*, 714–717.

Paulus, M.P., Geyer, M.A., and Braff, D.L. (1999). Long-range correlations in choice sequences of schizophrenic patients. Schizophr. Res. *35*, 69–75.

Paulus, M.P., Frank, L., Brown, G.G., and Braff, D.L. (2003a). Schizophrenia subjects show intact success-related neural activation but impaired uncertainty processing during decision-making. Neuropsychopharmacology *28*, 795–806.

Paulus, M.P., Hozack, N., Frank, L., Brown, G.G., and Schuckit, M.A. (2003b). Decision making by methamphetamine-dependent subjects is associated with error-rate-independent decrease in prefrontal and parietal activation. Biol. Psychiatry *53*, 65–74.

Paulus, M.P., Rogalsky, C., Simmons, A., Feinstein, J.S., and Stein, M.B. (2003c). Increased activation in the right insula during risk-taking decision making is related to harm avoidance and neuroticism. Neuroimage *19*, 1439–1448.

Paulus, M.P., Feinstein, J.S., Simmons, A., and Stein, M.B. (2004). Anterior cingulate activation in high trait anxious subjects is related to altered error processing during decision making. Biol. Psychiatry *55*, 1179–1187.

Paulus, M.P., Tapert, S.F., and Schuckit, M.A. (2005). Neural activation patterns of methamphetamine-dependent subjects during decision making predict relapse. Arch. Gen. Psychiatry *62*, 761–768.

Paulus, M.P., Lovero, K.L., Wittmann, M., and Leland, D.S. (2008). Reduced behavioral and neural activation in stimulant users to different error rates during decision making. Biol. Psychiatry *63*, 1054–1060.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R.J., and Frith, C.D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. Nature *442*, 1042–1045.

Platt, M.L., and Glimcher, P.W. (1999). Neural correlates of decision variables in parietal cortex. Nature *400*, 233–238.

Preuschoff, K., Bossaerts, P., and Quartz, S.R. (2006). Neural differentiation of expected reward and risk in human subcortical structures. Neuron *51*, 381–390.

Preuschoff, K., Quartz, S.R., and Bossaerts, P. (2008). Human insula activation reflects risk prediction errors as well as risk. J. Neurosci. *28*, 2745–2752.

Rangel, A., Camerer, C., and Montague, P.R. (2008). A framework for studying the neurobiology of value-based decision making. Nat. Rev. Neurosci. *9*, 545–556.

Ray, D., King-Casas, B., Montague, P.R., and Dayan, P. (2008). Bayesian Model of Behaviour in Economic Games. Adv. Neural Inf. Process. Syst. *21*, 1345–1353.

Rilling, J.K., Gutman, D.A., Zeh, T.R., Pagnoni, G., Berns, G.S., and Kilts, C.D. (2002). A neural basis for social cooperation. Neuron *35*, 395–405.

Rilling, J.K., Sanfey, A.G., Aronson, J.A., Nystrom, L.E., and Cohen, J.D. (2004). The neural correlates of theory of mind within interpersonal interactions. Neuroimage *22*, 1694–1703.

Rilling, J.K., Glenn, A.L., Jairam, M.R., Pagnoni, G., Goldsmith, D.R., Elfenbein, H.A., and Lilienfeld, S.O. (2007). Neural correlates of social cooperation and non-cooperation as a function of psychopathy. Biol. Psychiatry *61*, 1260–1271.

Roth, A.E. (1995). Introduction to Experimental Economics. In Handbook of Experimental Economics, J.H. Kagel and A.E. Roth, eds. (Princeton, NJ: Princeton University Press), pp. 3–109.

Roth, A.E. (2008). What have we learned from market design? Econ. J. *118*, 285–310.

Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E., and Cohen, J.D. (2003). The neural basis for economic decision-making in the ultimatum game. Science *300*, 1755–1758.

Saxe, R., Carey, S., and Kanwisher, N. (2004). Understanding other minds: linking developmental psychology and functional neuroimaging. Annu. Rev. Psychol. *55*, 87–124.

Schultz, W. (1998). Predictive reward signal of dopamine neurons. J. Neurophysiol. *80*, 1–27.

Schultz, W., and Dickinson, A. (2000). Neuronal coding of prediction errors. Annu. Rev. Neurosci. *23*, 473–500.

Schultz, W., Dayan, P., and Montague, P.R. (1997). A neural substrate of prediction and reward. Science *275*, 1593–1599.

Seger, C.A., Stone, M., and Keenan, J.P. (2004). Cortical Activations during judgments about the self and an other person. Neuropsychologia *42*, 1168–1177.

Seo, H., and Lee, D. (2007). Temporal filtering of reward signals in the dorsal anterior cingulate cortex during a mixed-strategy game. J. Neurosci. *27*, 8366–8377.

Seo, H., and Lee, D. (2009). Behavioral and neural changes after gains and losses of conditioned reinforcers. J. Neurosci. *29*, 3627–3641.

Seres, I., Unoka, Z., and Kéri, S. (2009). The broken trust and cooperation in borderline personality disorder. Neuroreport *20*, 388–392.

Singer, T., Seymour, B., O'Doherty, J.P., Stephan, K.E., Dolan, R.J., and Frith, C.D. (2006). Empathic neural responses are modulated by the perceived fairness of others. Nature *439*, 466–469.

Sripada, C.S., Angstadt, M., Banks, S., Nathan, P.J., Liberzon, I., and Phan, K.L. (2009). Functional neuroimaging of mentalizing during the trust game in social anxiety disorder. Neuroreport *20*, 984–989.

Sugrue, L.P., Corrado, G.S., and Newsome, W.T. (2004). Matching behavior and the representation of value in the parietal cortex. Science *304*, 1782–1787.

Sugrue, L.P., Corrado, G.S., and Newsome, W.T. (2005). Choosing the greater of two goods: neural currencies for valuation and decision making. Nat. Rev. Neurosci. *6*, 363–375.

Tom, S.M., Fox, C.R., Trepel, C., and Poldrack, R.A. (2007). The neural basis of loss aversion in decision-making under risk. Science *315*, 515–518.

Trivers, R.L. (1971). The evolution of reciprocal altruism. Q. Rev. Biol. *46*, 35–57.

Tse, W.S., and Bond, A.J. (2002a). Difference in serotonergic and noradrenergic regulation of human social behaviours. Psychopharmacology (Berl.) *159*, 216–221.

Tse, W.S., and Bond, A.J. (2002b). Serotonergic intervention affects both social dominance and affiliative behaviour. Psychopharmacology (Berl.) *161*, 324–330.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science *185*, 1124–1131.

Tversky, A., and Kahneman, D. (1982). Evidential impact of base rates. In Judgment under Uncertainty: Heuristics and Biases, D. Kahneman, P. Slovic, and A. Tversky, eds. (New York: Cambridge University Press), pp. 153–160.

Uddin, L.Q., Iacoboni, M., Lange, C., and Keenan, J.P. (2007). The self and social cognition: the role of cortical midline structures and mirror neurons. Trends Cogn. Sci. *11*, 153–157.

Unoka, Z., Seres, I., Aspán, N., Bódi, N., and Kéri, S. (2009). Trust game reveals restricted interpersonal transactions in patients with borderline personality disorder. J. Pers. Disord. *23*, 399–409.

van den Bos, W., Li, J., Lau, T., Maskin, E., Cohen, J.D., Montague, P.R., and McClure, S.M. (2008). The value of victory: social origins of the winner's curse in common value auctions. Judgm. Decis. Mak. *3*, 483–492.

Vogeley, K., and Fink, G.R. (2003). Neural correlates of the first-person-perspective. Trends Cogn. Sci. *7*, 38–42.

Vogeley, K., Bussfeld, P., Newen, A., Herrmann, S., Happe, F., Falkai, P., Maier, W., Shah, N.J., Fink, G.R., and Zilles, K. (2001). Mind reading: neural mechanisms of theory of mind and self-perspective. Neuroimage *14*, 170–181.

Vollenweider, F.X., Liechti, M.E., and Paulus, M.P. (2005). MDMA affects both error-rate dependent and independent aspects of decision-making in a two-choice prediction task. J. Psychopharmacol. *19*, 366–374.

Wallace, B., Cesarini, D., Lichtenstein, P., and Johannesson, M. (2007). Heritability of ultimatum game responder behavior. Proc. Natl. Acad. Sci. USA *104*, 15631–15634.

Wicker, B., Ruby, P., Royet, J.P., and Fonlupt, P. (2003). A relation between rest and the self in the brain? Brain Res. Brain Res. Rev. *43*, 224–230.

Wischniewski, J., Windmann, S., Juckel, G., and Brüne, M. (2009). Rules of social exchange: game theory, individual differences and psychopathology. Neurosci. Biobehav. Rev. *33*, 305–313.

Wood, R.M., Rilling, J.K., Sanfey, A.G., Bhagwagar, Z., and Rogers, R.D. (2006). Effects of tryptophan depletion on the performance of an iterated Prisoner's Dilemma game in healthy adults. Neuropsychopharmacology *31*, 1075–1084.

Yi, R., Buchhalter, A.R., Gatchalian, K.M., and Bickel, W.K. (2007). The relationship between temporal discounting and the prisoner's dilemma game in intranasal abusers of prescription opioids. Drug Alcohol Depend. *87*, 94–97.

Yoshida, W., Seymour, B., Friston, K.J., and Dolan, R.J. (2010). Neural mechanisms of belief inference during cooperative games. J. Neurosci. *30*, 10744–10751.