not expect the human psyche to have evolved to deal with one-shot games. Under normal circumstances (and especially before the dawn of the global village), one can rarely (if ever) be absolutely certain that one will never interact with the same person again. Even if two individuals never interact again, others may observe the interaction (Alexander 1987). For this reason, humans may have internal rewards for acting cooperatively in repeated interactions, which evolved (or were learned) because those rewards caused people to cooperate and reap the benefits of mutual cooperation. These internal rewards (or nonstandard preferences such as a positive valuation of fairness, equity, or the well-being of individual exchange partners) would also cause them to act cooperatively in the novel case of one-shot interactions as well.

The target article's more contentious claim is that game theoretical rationality cannot be salvaged as a model of human decision-making in social situations by incorporating nonstandard preferences into the decision makers' utility functions. In section 8.1, Colman illustrates the problem of finding compromise solutions where individual preferences differ with Sugden's example of a family going for a walk, and asserts that tinkering with utility functions cannot explain their solubility. He insists that the "team reasoning" by which compromises are negotiated is an "inherently non-individualistic" process. However, we looked in vain for evidence or argument in support of these conclusions. It is, after all, individuals who ultimately make the choices in experimental games, so if "a team reasoning player" really seeks to maximize "joint or collective payoff," as Colman claims (sect. 8.1), then contra his own conclusion (sect. 9.1), this is evidence of nonstandard preferences, not of "nonstandard types of reasoning." Moreover, such a process of team reasoning cannot have general applicability to social dilemmas with divisible payoffs, because it is inconsistent with the evidence that experimental subjects will pay to punish other players (Fehr & Gächter 2000; Roth 1995). We do not understand the notion of a "psychological" theory that is "non-individualistic"; the individual organism is psychology's focal level of analysis.

We agree with Colman in saying that game theoretic rationality does not accurately describe human social behavior. However, he has not argued convincingly why expanding calculations of Expected Utility to include nonstandard preferences and rational responses to irrational behavior cannot salvage models of Expected Utility, so we would argue that such expanded models still may be effective at explaining human behavior. Evolutionary models can help generate hypotheses about what those nonstandard preferences are, and how we might expect people to respond to apparently irrational behavior.

ACKNOWLEDGMENTS

We would like to thank L. DeBruine and M. Wilson.

Neural game theory and the search for rational agents in the brain

Gregory S. Berns

Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, Atlanta, GA 30322. **gberns@emory.edu**
http://www.ccnl.emory.edu

Abstract: The advent of functional brain imaging has revolutionized the ability to understand the biological mechanisms underlying decision-making. Although it has been amply demonstrated that assumptions of rationality often break down in experimental games, there has not been an overarching theory of why this happens. I describe recent advances in functional brain imaging and suggest a framework for considering the function of the human reward system as a discrete agent.

The assumption of rationality has been under attack from several fronts for a number of years. Colman succinctly surveys the evidence against rationality in such ubiquitous decision games ranging from the Prisoner's Dilemma (PD) to Centipede and comes to the conclusion that standard game theory fails to explain much of human behavior, especially within the confines of social interactions. This is a startling conclusion, and not merely because the tools of game theory have become so enmeshed as an approach to decision-making and risk management. The startling implication is that almost every commonplace decision that humans make is socially constrained. Whether it is an explicit social interaction like the PD or an implicit social construct that underlies the decision to stay a few more hours at work versus spending time with family, social connectedness cannot be factored out of almost any meaningful human decision. If the assumption of rationality governs the application of game theoretic tools to understanding human decision-making, then its very failure in social domains brings into question its practical utility. Can the tools of game theory be applied within ad hoc frameworks like behavioral game theory, or psychological game theory? The reunification of psychological principles within economics is a necessary first step (Camerer 1999). However, like cognitive psychology before it, psychological principles also often fail to explain human behavior. Neuroscience offers yet another perspective on human behavior, and the application within economic frameworks has come to be called neural economics (Montague & Berns 2002; Wilson 1998).

One of the earliest applications of functional brain imaging, in this case functional magnetic resonance imaging (fMRI), to the evaluation of the neural basis of game theory was performed by McCabe and colleagues (McCabe et al. 2001). In a variant of Centipede, pairs of subjects played three types of games (trust, punish, and mutual advantage). As Colman points out, most players behave cooperatively in these types of games – an observation unexplainable by standard rational game theory. McCabe's results implicated a specific region of the medial prefrontal cortex that subserved this type of cooperative behavior. Perhaps because of the small sample size ($N = 6$), statistically significant results were not found for the noncooperators (i.e., "rational" players), and nothing could be said about the possible existence of rational agents in the brain.

In a recent study, our group used fMRI to examine the neural responses of one player in an all-female PD interaction (Rilling et al. 2002). The large sample size ($N = 36$) yielded reasonable power to detect a number of significant activations related to the different outcomes. At the simplest level, striatal activation was most strongly associated with mutual cooperation outcomes. When subjects played a computer, the striatal activation was greatly reduced, suggesting that the striatal activity was specifically modulated by the presence or absence of social context. This region of the striatum is of particular interest because it is the same region most closely associated with hedonic reward processes (Schultz et al. 1997). Activation of the ventral striatum, especially the nucleus accumbens, has been observed repeatedly in various forms of appetitive Pavlovian conditioning and drug administration – activity that is widely believed to be modulated by dopamine release (Robbins & Everitt 1992). The striatal activation observed with mutual cooperation most likely reflected the overall utility of that outcome in the context of the PD. The same region of striatum was also observed to be active during the decision-making phase of the experiment, but only when the subject chose to cooperate following her partner's cooperation in the previous round. This latter finding suggests that the striatum was not only encoding the actual utility of the outcome, but the expected utility during the decision-making phase. We do not yet know the exact relationship between reward-system activity and expected utility (modified or not), but the mesolimbic reward system appears to be a promising candidate for a "rational agent" within the brain.

Based on the results of the PD experiment, our group realized that it would be desirable to monitor brain activity in both players simultaneously. The rationale is that by monitoring the activity in the reward pathways of both players in a two-player game, one should have a direct assay of the player's expected utility functions

without resorting to revealed preference. Under rational assumptions, these functions should be about the same. In a proof of principle experiment, based loosely on the game of matching pennies, we described the methodology necessary to conduct such simultaneous imaging experiments, which we have termed, "Hyperscanning" (Montague et al. 2002). In this first experiment, we did not undertake an assessment of utility functions, but it is worth pointing out that the methodology is generalizable to *N*-person interactions.

There is good reason to be hopeful that neuroscientific methods, especially functional brain imaging, will help resolve the apparent paradoxes between rational game theory and both behavioral and psychological variants. By looking inside the brain, it becomes possible to identify specific neuronal clusters that may be operating near different equilibria. Recent work suggests that neurons in the lateral intraparietal area encode expected utilities during probabilistic reward paradigms in monkeys (Glimcher 2002; Gold & Shadlen 2001). In humans, correlates of utility, as predicted by prospect theory, have been found in discrete elements of the reward pathway (Breiter et al. 2001). Taken together, these early neuroscientific enquiries suggest that game theoretic principles are very much viable predictors of neuronal behavior. The interaction of different pools of neurons in the brain may result in phenotypic behavior that appears to be irrational, but it is possible that the rational agents are the neurons, not the person.

ACKNOWLEDGMENTS

The author is supported by grants from the National Institutes of Health (K08 DA00367, RO1 MH61010, and R21 DA14883).

Evolution, the emotions, and rationality in social interaction

David J. Butler

Department of Economics, University of Arizona, Tucson, AZ, 85721 and Department of Economics, University of Western Australia, Nedlands, WA 6009, Australia. **dbutler@eller.arizona.edu**

Abstract: Although Colman's criticisms of orthodox game theory are convincing, his assessment of progress toward construction of an alternative is unnecessarily restrictive and pessimistic. He omits an important multidisciplinary literature grounded in human evolutionary biology, in particular the existence and function of social emotions experienced when facing some strategic choices. I end with an alternative suggestion for modifying orthodox game theory.

Colman has brought together an impressive collection of arguments to demonstrate both serious weaknesses and failures of orthodox game-theoretic rationality. But to address these problems he offers only some "tentative and ad hoc suggestions" (sect. 8, para. 2) from psychological game theory. Although I strongly endorse his criticisms of orthodox game theory and agree that the new reasoning principles he describes have a part to play, I think his discussion of "where next" neglects some important ideas from a recent and exciting multidisciplinary literature.

Because of the newness of this research and its multidisciplinary origins, we must piece together some apparently disparate strands of thought in order to glimpse the beginnings of an alternative to orthodox game-theoretic rationality. One reason why Colman's "destruction" work is much more comprehensive and convincing than his subsequent "construction" is his early distinction between the nonrational "mindless" (sect. 1.3, para. 3) strategic interaction of evolutionary game theory, and the rational strategic interaction of human agents. He argues the former is not of interest for his views on rationality, but I will argue that this dichotomy severely restricts the variety of new ideas that he can consider.

To understand human decision-making in social interactions, we should keep in mind that both humans and their decision-making apparatus are themselves products of natural selection. There is a growing consensus behind the "social animal" hypothesis (e.g., Barton & Dunbar 1997; Cartwright 2000), which maintains that the selection pressures among humans were primarily an intraspecies phenomenon. In successive generations, reproductive success went to those with the best grasp of the complexities of the "social chess" that was a constant theme of tribal life. In this view, underlying the anomalous cooperation observed in both experimental and real world social dilemmas is an innate predisposition, not for unconditional cooperation, but for some form of reciprocity. Indeed, there is a now a significant literature in experimental and theoretical economics on reciprocity models (see Sethi & Somanathan 2003 for a recent survey).

Trivers (1985) argued that reciprocal altruism in humans evolved by molding our emotional responses to the cost/benefit calculus of social exchange; among these emotions are both cooperative and punitive sentiments. In a recent study, Price et al. (2002) demonstrate that "punitive sentiments in collective action contexts have evolved to reverse the fitness advantages that accrue to free riders over producers." Indeed, punitive sentiments must go hand in hand with a preparedness to risk cooperation if cooperation is to survive the process of natural selection.

There is also a growing recognition that contrary to the standard model of rational choice, "gut feelings experienced at the moment of making a decision, which are often quite independent of the consequences of the decision, can play a critical role in the choice one eventually makes" (Loewenstein et al. 2001). For example, they refer to the work of the neuroscientist Damasio (1994), who shows how our ability to choose rationally is intertwined with our ability to experience emotional reactions to the choices we face. Damasio calls these reactions "somatic markers" and argues: "Nature appears to have built the apparatus of rationality (the cerebral cortex) not just on top of the apparatus of biological regulation (the limbic system), but also from it and with it" (p. 128). A more human rationality may also allow for heterogeneity of choices, in recognition of the differing intensities with which the social (and other) emotions are experienced by different people in the deliberation process.

Although neither Damasio nor Loewenstein and colleagues directly address the social emotions, we can easily extend their arguments to the context of strategic interaction, where the emotions that need incorporating for a descriptive theory are the cooperative and punitive sentiments behind reciprocity. We might even go further and argue for their incorporation into normative models, as well. This is because our emotional responses to choices that place our individual and collective interests in opposition embody adaptive knowledge that helped win many games of "social chess." These somatic responses may help us to extract the long run benefits of cooperation.

There is also now direct evidence that a somatic response specific to human strategic interactions exists. Recent work by Rilling et al. (2002), using fMRI scans on subjects playing prisoner's dilemma games, found that an individual's brain activation patterns when the playing partner was identified as a human differ from when the partner was identified as a computer. They conclude "that (the relevant activation patterns) may relate specifically to cooperative social interactions with human partners." It seems that human players rely more on a common knowledge of *humanity* in strategic interaction than a common knowledge of *rationality* as conventionally understood.

The finding of Rilling and colleagues also highlights the importance of the description or "framing" of the game for our choices. Loewenstein and colleagues also noted, for choice under risk, that these factors become important when we incorporate emotions experienced when choosing, in contrast to the purely cognitive evaluations of the standard model that are supposedly context independent. This implies we can no longer expect game theoretic models to satisfy description invariance if a change in the description (e.g., that the other player is a person or a program) is implemented.