sounds, though perhaps a uniquely human capability, mind-reading is the ability to reason about the otherwise unobservable mental states of others and make predictions about their behaviors based partly on the awareness that others are intentional agents with general goals similar to one's own. Colman uses the phrasing of mind-reading in his description of how a Stackelberg-reasoning player might deliberate.

It seems that cultural mechanisms solve cooperative problems so transparently, however, that many do not recognize them as solutions at all. Broadly speaking, communicating via spoken language can be construed as mind-reading. I can utter a sound and others can predict my intent based on that sound, unless they do not share my otherwise arbitrary association between sound and meaning. Part of the problem of classic analytic game theory revolves around the standard assumption that players do not speak to one another. This assumption is put into practice in experimental game research where subjects usually do not communicate during experiments. It seems that pre-game communication among subjects is such a simple solution to many games that researchers routinely disallow it in order for the "truly" interesting solutions to emerge (van Huyck et al. 1990). Although "cheap talk" solutions may seem trivial to game theoreticians, because all extant humans can easily communicate this way, from a comparative evolutionary perspective, such a solution is far from trivial. Although simple communication among players is often sufficient to generate complexly coordinated behaviors, speaking is anything but simple. Such a research design excludes the very solution that natural selection likely created to solve the problem. Experimental social games in which subjects are not allowed to speak to one another are a bit like sports competitions where subjects must compete with their legs shackled together.

Verbalizing intent may be feasible in small groups, but how do humans communicate expectations between members of large cooperative groups like those that characterize most human societies – ethnic groups, for example – in which many interactions are seemingly anonymous? How do fellows know that they share beliefs concerning behavior critical for coordination? How can individuals predict what others think and will do in such large groups? There are a number of options. One could attempt to learn, on one's own, the beliefs of all the potential cooperative partners. This could prove difficult, time-consuming, and error-prone (Boyd & Richerson 1995). In addition to speaking, however, humans use symbols and markers of group identity to transmit information that helps them make predictions about the otherwise unobservable mental states of others. McElreath et al. (2003) argue that group markers, such as speech or dress, function to allow individuals to advertise their behavioral intent so that individuals who share social norms can identify one another and assort for collective action. Although cheaters are a problem if interaction is structured like a prisoner's dilemma, McElreath et al.'s critical point is that group markers are useful if people engage in social interactions structured as coordination games. Colman notes the advantages of making predictions about the behavior of others based on information acquired culturally.

The great potential payoffs from successfully navigating real-life coordination games may have been part of the selective pressure favoring the evolution of language and culture. Coordination problems abound, and their solutions are facilitated when players have the ability to quickly acquire expectations about fellow players' behavior. Whether such adaptations are rational or not, ignoring the evolutionary mechanisms that produced these cognitive abilities is a mistake.

Humans should be individualistic and utility-maximizing, but not necessarily "rational"

Pat Barclay and Martin Daly

Department of Psychology, McMaster University, Hamilton, Ontario, L8S 4K1 Canada. **barclapj@mcmaster.ca** **daly@mcmaster.ca**
http://www.science.mcmaster.ca/Psychology/md.html

Abstract: One reason why humans don't behave according to standard game theoretical rationality is because it's not realistic to assume that everyone else is behaving rationally. An individual is expected to have psychological mechanisms that function to maximize his/her long-term payoffs in a world of potentially "irrational" individuals. Psychological decision theory has to be individualistic because individuals make decisions, not groups.

Game theoretical rationality in the service of personal profit maximization is not an adequate model of human decision-making in social bargaining situations. This proposition is a large part of Colman's thesis, and we have no quarrel with it. Does anyone? The point is proven whenever experimental subjects reject offers in Ultimatum Games, share the pot in Dictator Games, or cooperate in one-shot Prisoner's Dilemmas (e.g., Frank et al. 1993; Roth 1995). The idea that this simple game theoretical account is descriptive rather than normative is surely dead in experimental economics and psychology. Evolutionary models are an important exception, because they purport to describe what strategies will be selected for. However, in these models, the concept of "rationality" is superfluous, because the selection of superior strategies occurs by a mindless, competitive process (Gintis 2000).

One way in which rational choice theory (RCT) is problematic is the default expectation that all other players will behave "rationally." Individuals can be expected to occasionally make decisions that are not in accordance with predictions of RCT because of incomplete information, errors, concern for the welfare of others (such as friends or relatives), or manipulation by others. Also, individuals may be expected to act irrationally if that irrationality is more adaptive than rationality. For example, Nowak et al. (2000) show that the "irrational" behavior of demanding fair offers in the Ultimatum Game is evolutionarily stable if each individual has knowledge about what kind of offers each other individual will accept. Similarly, aggressive behavior or punishment, while not "rational" in the game theoretic sense, can evolve if the costs of being punished are high (Boyd & Richerson 1992), because the punished individual learns (potentially via operant conditioning) to desist from the behavior that brought on the punishment.

Given that others are sometimes not strictly rational, an instrumentally rational individual should reevaluate his/her situation and act accordingly (Colman hints at this in sect. 8.4). We argue that rationality should not even be the default assumption because individuals are repeatedly faced with evidence (from real life) that others are not always rational, and this affects the strategy that a profit-maximizing individual should take. For example, when playing an iterated Prisoner's Dilemma against what appears to be a conditional cooperator (such as a Tit-for-Tat player), a rational and selfish player should cooperate for a while. Even if the rational actor is cheated in later rounds, he/she has still done better than if he/she had never cooperated. Thus, a rational player should attempt to determine the likely responses of others, rather than assume that (despite past experience to the contrary) they will be "rational." Henrich et al. (2001) argue that when people play experimental games, they compare the games to analogous situations with which they have experience. If different people have had different experiences because of different backgrounds, then they will have different beliefs about how others will behave. Thus, in iterated games, each player may be acting rationally with respect to his/her past experience. Recent experiences have large effects on how people play experimental games (Eckel & Wilson 1998a), possibly because players use their experience in the games to update their beliefs of what others will do.

This does not explain behavior in one-shot games, but we would 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