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.

https://doi.org/10.1017/S0140525X03250056 Published online by Cambridge University Press

Colman discusses "Stackelberg reasoning" and "team thinking," and he mentions (sect. 8.1, para. 3) that the collective preferences of team reasoning can be triggered by the acceptance of a group identity in certain contexts. But he doesn't explain where these alternative reasoning methods come from, how they survive, or how, if cooperation in social dilemmas is sensitive to the cost/benefit ratio, we might "trade-off" the different reasoning methods in some meta-reasoning process. Hamilton's (1964) "kin-selection," Trivers' (1971) "reciprocal altruism," and Alexander's (1987) "indirect reciprocity" models might at least offer a way to think about answering these questions.

If we wish to incorporate the social emotions triggered by a strategic choice into our models, how might we proceed? Hollis and Sugden (1993) explained (p. 28) why our attitudes toward consequences cannot be simply "bundled in" with the existing utilities of a game. A more plausible path then may be to alter the weighting we attach to the consequences, along the lines of the "rank dependent" transformation of the cumulative probability distribution, which has worked so well among the alternatives to expected utility theory (see Starmer 2000). In this way, some plausible improvements to orthodox game theory might be developed, as has already happened to expected utility theory in choice under risk.

Behavioral game theory: Plausible formal models that predict accurately

Colin F. Camerer

Division of Social Sciences, California Institute of Technology, Pasadena, CA 91125. **camerer@hss.caltech.edu** **http://hss.caltech.edu/~camerer**

Abstract: Many weaknesses of game theory are cured by new models that embody simple cognitive principles, while maintaining the formalism and generality that makes game theory useful. Social preference models can generate team reasoning by combining reciprocation and correlated equilibrium. Models of limited iterated thinking explain data better than equilibrium models do; and they self-repair problems of implausibility and multiplicity of equilibria.

Andrew Colman's wonderful, timely, and provocative article collects several long-standing complaints about game theory. Part of the problem is that game theory has used lots of applied math and little empirical observation. Theorists think that deriving perfectly precise analytical predictions about what people will do (under differing assumptions about rationality) from pure reasoning is the greatest challenge. Perhaps it is; but why is this the main activity? The important uses of game theory are prescriptive (e.g., giving people good advice) and descriptive (predicting what is likely to happen), because good advice (and good design of institutions) requires a good model of how people are likely to play. It is often said that studying analytical game theory helps a player understand what might happen, vaguely, even if it does not yield direct advice. This is like saying that studying physics helps you win at pool because the balls move according to physical laws. A little physics probably doesn't hurt, but also helps very little compared to watching other pool players, practicing, getting coaching, studying what makes other players crumble under pressure, and so on.

While Colman emphasizes the shortcomings of standard theory, the real challenge is in creating new theory that is psychological (his term) or "behavioral" (my earlier term from 1990; they are synonymous). Models that are cognitively plausible, explain data (mostly experimental), and are as general as analytical models, have developed very rapidly in just the last few years. Colman mentions some. Others are described in my book (Camerer 2003).

An important step is to remember that games are defined over utilities, but in the world (and even the lab) we can usually only measure pecuniary payoffs – status, territory, number of offspring, money, and so forth. The fact that people cooperate in the prisoner's dilemma (PD) is *not* a refutation of game theory per se; it is a refutation of the joint hypothesis of optimization (obeying dominance) and the auxiliary hypothesis that they care only about the payoffs we observe them to earn (their own money). The self-interest hypothesis is what's at fault.

Several new approaches to modeling this sort of "social preferences" improve on similar work by social psychologists (mentioned in sect. 8.1), because the new models are designed to work across games and endogenize when players help or hurt others. For example, in Rabin's fairness theory, player A treats another player's move as giving herself (A) a good or bad payoff, and forms a judgment of whether the other player is being nice or mean. Players are assumed to reciprocate niceness and also meanness.

Rabin's model is a way to formalize *conditional* cooperation – people cooperate if they expect others to do so. This provides a way to anchor the idea of "team reasoning" in methodological individualism. In experiments on group identity and cooperation, a treatment (like subjecting subjects to a common fate or dividing them into two rooms) or categorization (whether they like cats or dogs better) is used to divide subjects into groups. In the Rabin approach, PD and public goods games are coordination games in which players are trying to coordinate on their level of mutual niceness or meanness.

Experimental identity manipulations can be seen as correlating devices that tell subjects which equilibrium will be played, that is, whether they can expect cooperation from the other players or not (which is self-enforcing if they like to reciprocate). This explanation is *not* merely relabeling the phenomenon, because it makes a sharp prediction: A correlated equilibrium requires a publicly observable variable that players commonly know. If identity is a correlating device, then when it is not commonly known, cooperation will fall apart. For example, suppose members of the A team ("informed A's") are informed that they will play other A's, but the informed As' partners will not know whether they are playing A's or B's. Some theories of pure empathy or group identification predict that who the other players think they are playing won't matter to the informed A's because they just like to help their teammates. The correlated equilibrium interpretation predicts that cooperation will shrink if informed A's know that their partners don't know who they are playing, because A's only cooperate with other A's *if they can expect cooperation by their partners.* So there is not necessarily a conflict between an individualist approach and team reasoning: "Teamness" can arise purely through the conjunction of reciprocal individual preferences and observable correlating variables, which create shared beliefs about what team members are likely to do. What those variables are is an interesting empirical matter.

Another type of model weakens the mutual consistency of players' choices and beliefs. This might seem like a step backward but it is not – in fact, it solves several problems that mutual consistency (equilibrium) *creates.* In the cognitive hierarchy (CH) model of Camerer et al. (2002), a Poisson distribution of discrete levels of thinking is derived from a reduced-form constraint on working memory. Players who use 0 levels will randomize. Players at level $K > 0$ believe others are using 0 to K to 1 levels. They know the normalized distribution of lower-level thinkers, and what those others do, and best respond according to their beliefs. The model has one parameter, τ, the average number of levels of thinking (it averages around 1.5 in about a hundred games). In the CH model, every strategy is played with positive probability, so there are no incredible threats and odd beliefs after surprising moves. Once τ is fixed (say 1.5), the model produces an exact statistical distribution of strategy frequencies – so it is *more* precise in games with multiple equilibria, and is generally *more* empirically accurate than equilibrium models. The model can explain focal points in matching games if level-0 subjects choose what springs up. The model also has "economic value": If subjects had used it to forecast what others were likely to do, and best responded to the model's advice, they would have earned substantially more (about

https://doi.org/10.1017/S0140525X03250056 Published online by Cambridge University Press