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.

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.