Beliefs, intentions, and evolution: Old versus new psychological game theory

Jeffrey P. Carpenter and Peter Hans Matthews

Economics Department, Middlebury College, Middlebury, VT 05753.

jpc@middlebury.edu  peter.h.mathews@middlebury.edu

http://community.middlebury.edu/~jcarpent/  http://community.middlebury.edu/~pmathew/

Abstract: We compare Colman’s proposed “psychological game theory” with the existing literature on psychological games (Geanakoplos et al. 1989), in which beliefs and intentions assume a prominent role. We also discuss experimental evidence on intentions, with a particular emphasis on reciprocal behavior, as well as recent efforts to show that such behavior is consistent with social evolution.

Andrew Colman’s target article is a call to build a new, psychological, game theory based on “nonstandard assumptions.” Our immediate purpose is to remind readers that the earlier work of Geanakoplos et al. (1989), henceforth abbreviated as GPS, which the target article cites but does not discuss in detail, established the foundations for a theory of “psychological games” that achieves at least some of the same ends. Our brief review of GPS and some of its descendants — in particular, the work of Rabin (1993) and Falk and Fischbacher (2000) — will also allow us to elaborate on the connections between psychological games, experimental economics, and social evolution.

The basic premise of GPS is that payoffs are sometimes a function of both actions and beliefs about these actions, where the latter assumes the form of a subjective probability measure over the product of strategy spaces. If these beliefs are “coherent” — that is, if they conform to Bayes’ rules — and so on — and this coherence is common knowledge, then the influence of second (and higher) order beliefs can be reduced to a set of common first-order beliefs. That is, in a two-player psychological game, for example, the utilities of A and B are functions of the strategies of each and the beliefs of each about these strategies. A psychological Nash equilibrium (PNE) is then a strategy profile in which, given their beliefs, neither A nor B would prefer to deviate, and these first-order beliefs are correct. If these augmented utilities are continuous, then augmented utilities can be derived from a “material game” with the addition of parsimony. As such, it appears that intentions matter to decision-makers: consider the legal difference between manslaughter and murder — and that game theorists would do well to heed the advice of Colman and others who advocate a more behavioral approach.

For a time, it was not clear whether or not the GPS framework was tractable. Rabin (1993), which Colman cites as an example of behavioral, rather than psychological, game theory, was perhaps the first to illustrate how a normal form psychological game could be derived from a “material game” with the addition of parsimonious “kindness beliefs.” In the standard two-person prisoner’s dilemma (PD), for example, he showed that the “all cooperate” and “all defect” outcomes could both be rationalized as PNEs. It is obvious that intentions matter to decision-makers: consider the legal difference between manslaughter and murder — and that game theorists would do well to heed the advice of Colman and others who advocate a more behavioral approach.

This result does not tell us, though, whether this outcome is consistent with the development of reciprocal intentions or norms over time, or, in other words, whether social evolution favors those with “good intentions.” To be more concrete, suppose that the proposers and responders in the ultimatum game are drawn from two distinct populations and matched at random each period, and that these populations are heterogeneous with respect to intention. Could these intentions survive “selection” based on differences in material outcomes? Or do these intentions impose substantial costs on those who have them?

There are still no definitive answers to these questions, but the results in Binmore et al. (1988), henceforth abbreviated as BCS, hint that prosocial intentions often survive. BCS consider a “miniature ultimatum game” with a limited strategy space and show there are two stable equilibria within this framework. The first corresponds to the subgame perfect equilibrium — proposers are selfish, and responders accept these selfish offers — but in the second, proposers are fair and a substantial share of responders would turn down an unfair offer. Furthermore, these dy-
To have and to eat cake: The descriptive role of game-theoretic explanations of human choice behavior

William D. Casebeer* and James E. Parco*
*Department of Philosophy, United States Air Force Academy, Colorado Springs, CO 80840; b American Embassy, Tel Aviv, 63903 Israel.
http://www.usafa.af.mil/dfpra/CVs/Casebeer.html
http://parco.usafa.bsz

Abstract: Game-theoretic explanations of behavior need supplementation to be descriptive; behavior has multiple causes, only some governed by traditional rationality. An evolutionarily informed theory of action cannot simply overlook the overlapping causal domains: neurobiological, psychological, and rational. Casebeer's discussion is insufficient because he neither evaluates existing learning models nor qualifies under what conditions his propositions hold. Still, inability to incorporate emotions in axiomatic models highlights the need for a comprehensive theory of functional rationality.

The power and beauty of von Neumann and Morgenstern's Theory of Games and Economic Behavior (1944) and Luce and Raiffa's Games and Decisions (1957) lie in their mathematical coherence and axiomatic treatment of human behavior. Once rational agents could be described mathematically, game theory provided a far-reaching normative model of behavior requiring an assumption of common knowledge of rationality. This assumption (in addition to the often unstated requirement that a player fully understand the game situation) is subsumed under the phrase "the theory assumes rational players" (Luce & Raiffa 1957). But we know that, descriptively speaking, this is not always the case. The literature has clearly shown that not only are these (mathematically required) assumptions often too strong to be met in practice, but also that the "rational actor theory" (hereafter RAT) is underspecified in that it cannot effectively accommodate emotions. But does this constitute a failure of RAT? We think not.

Nevertheless, we agree with Colman's larger point that we need a "psychological game theory," or rather, a neurobiologically informed theory of decision-making. This is not because of the spectacular failure of game theoretic assumptions in any particular experiment, but rather stems from an ecumenically and fundamentally naturalizable worldview about the causes of, and norms governing, human behavior. Choice-driven behavior is a function of multiple, highly distributed brain subsystems that include affect and emotion. For example, in the domain of moral judgment, good moral cognition is driven by a variety of brain structures, only some involved in ratiocination as traditionally construed (Casebeer & Churchland 2003). Even the most ardent RAT enthusiast recognizes that if your explanation is all human behavior, your explanations will be more comprehensive than adhering to RAT alone.

Thus, we question the usefulness of Colman's ad hoc refinements for prescriptions of behavior in interactive decision-making, primarily because he has neither (1) qualified his theory as to when and under what conditions it applies, nor (2) provided an account for learning in games (beyond simple Stackelberg reasoning). For example, Colman uses the two-player centipede game as a primary domain in which he justifies his theory. However, recent evidence experimentally investigating three-player centipede games (Parco et al. 2002) directly contradicts it. Parco et al. extended the McKelvey and Palfrey (1992) study to three players using small incentives (10 cents for stopping the game at the first node, and $25.60 for continuing the game all the way to the end) and obtained similar results, soundly rejecting the normative equilibrium solution derived by backward induction. However, when the payoffs of the game were increased by a factor of 50 (and each player had the opportunity to earn $7,680), the results were markedly different. Although initial behavior of both the low-pay and high-pay conditions mirrored that of the McKelvey and Palfrey study, over the course of play for 60 trials, behavior in the high-pay treatment converged toward the Nash equilibrium and could be well accounted for using an adaptive reinforcement-based learning model. Furthermore, as noted by McKelvey and Palfrey (1992) and later by Fey et al. (1996), in all of the centipede experiments that were conducted until then, there were learning effects in the direction of equilibrium play. Colman's oversight of the extent learning in games literature and his brief account for the dynamics of play through Stackelberg reasoning is insufficient.

Learning in games manifests itself in a variety of processes quite different from simple Stackelberg reasoning (see Camerer & Ho 1999, Erev & Roth, 1998). For example, Rapoport et al. (2002) document almost "magical" convergence to the mixed-strategy equilibrium over 70 trials without common knowledge or between-trial feedback provided to subjects. Neither traditional game theory nor Colman's model can account for such data.

Generally speaking, Colman does little to improve prescriptive recommendations for human behavior both within and outside of the subset of games he has described; his paper is really a call for more theory than a theory proper. RAT's difficulty in dealing with emotions serves as proof-of-concept that we need a more comprehensive theory. Humans are evolved creatures with multiple causes of behavior, and the brain structures that subserve "rational" thought are, on an evolutionary timescale, relatively recent arrivals compared to the midbrain and limbic systems, which are the neural mechanisms of affect and emotion. Ultimately, our goal should be to formulate an explanation of human behavior that leverages RAT in the multiple domains where it is successful, but that also enlightens (in a principled way) to when and why RAT fails. This more comprehensive explanation will be a neurobiological cum psychological cum rational theory of human behavior.

The problems game-theoretic treatments have in dealing with the role of emotions in decision-making serve to underscore our point. There are at least two strategies "friends of RAT" can pursue: (1) attempt to include emotions in the subjective utility function (meaning you must have a mathematically rigorous theory of the emotions, this is problematic), or (2) abandon RAT's claim to be discussing proximate human psychology and, instead, talk about how emotions fit in system-wide considerations about long-term strategic utility (Frank 1988). The latter approach has been most successful, although it leaves RAT in the position of being a distal explanatory mechanism. The proximate causes of behavior in this story will be locally antithetical or possibly irrational (hence the concerns with emotions). How would "new wave RAT" deal with this? One contender for a meta-theory of rationality that can accommodate the explanatory successes of RAT, yet can also cope with their failure in certain domains, is a functional conception of rationality. The norms that govern action are reasonable, and reason-giving for creatures that wish to be rational, insofar as such norms allow us to function appropriately given our evolutionary history and our current environment of action (Casebeer 2003).

We acknowledge that RAT will require supplementation if it is to fully realize its descriptive explanatory role of predicting human action and providing us with a normative yardstick for it. Utility theory must incorporate neurobiological and psychological deter-