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 a third of the economic value of perfect advice). Nash equilibrium, in contrast, sometimes has negative economic value.

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.matthews@middlebury.edu
http://community.middlebury.edu/~jcarpent/
http://community.middlebury.edu/~pmatthew/

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, the information embodied in second-order beliefs are consistent with the first-order beliefs, 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 all normal form psychological games must have at least one PNE.

The introduction of beliefs provides a natural framework for modeling the role of intentions in strategic contests, and this could well prove to be the most important application of GPS. 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.

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.

As Rabin (1993) himself notes, this transformation of the PD is not equivalent to the substitution of altruistic agents for self-interested ones: the "all defect" outcome, in which each prisoner believes that the other(s) will defect, could not otherwise be an equilibrium. This is an important caveat to the recommendation that we endow economic actors with "nonstandard reasoning processes," and prompts the question: What observed behavior will the "new psychological game theory" explain that an old(er) GPS-inspired one cannot? Or, in narrower terms, what are the shortcomings of game theoretic models that incorporate the role of intentions, and therefore such emotions as surprise or resentfulness?

The answers are not obvious, not least because there are so few examples of the transformation of material games into plausible psychological ones, and almost all of these share Rabin's (1993) emphasis on kindness and reciprocal behavior. It does seem to us, however, that to the extent that Colman's "nonstandard reasoning" can be formalized in terms of intentions and beliefs, there are fewer differences between the old and new psychological game theories than at first it seems.

There is considerable experimental evidence that intentions matter. Consider, for example, Falk et al. (2000), in which a first mover can either give money to, or take money away from, a second mover, and any money given is tripled before it reaches the second mover, who must then decide whether to give money back, or take money from, the first mover. Their analysis suggests that there is a strong relationship between what the first and second movers do: in particular, *the more the first mover gives (takes), the more the second mover takes (gives) back.*

Falk et al. (2000) find that first mover giving (taking) is interpreted as a friendly (unfriendly) act, and that these intentions matter. Without the influence of beliefs or intentions on utilities, there would be a single Nash equilibrium in which the first mover takes as much as possible because she "knows" that the second has no material incentive to retaliate. Although this behavior can also be supported as a PNE, so can that in which the first mover gives and expects a return and the second mover understands this intention and reciprocates. When the experiment is changed so that the first mover's choice is determined randomly, and there are no intentions for the second mover to impute, the correlation between first and second mover actions collapses. We see this as evidence that beliefs – in particular, intentions – matter, but also that once these beliefs have been incorporated, a modified "rational choice framework" is still useful.

Building on both GPS and Rabin (1993), Dufwenberg and Kirchsteiger (1998) and Falk and Fischbacher (2000) derive variations of Rabin's (1993) "fairness equilibrium" for extensive form games, with results that are consistent with experimental evidence. The simplest of these is the ultimatum game, in which a first mover offers some share of a pie to a second mover who must then accept or reject the proposal. With kindness functions similar to Rabin's (1993), Falk and Fischbacher (2000) show that the ultimatum game has a unique PNE that varies with the "reciprocity parameters" of proposer and responder. Furthermore, this equilibrium is consistent with the observations that the modal offer is half the surplus, that offers near the mode are seldom rejected, that there are few of the low offers that are consistent with the subgame perfect equilibrium, and that most of these low offers are rejected.

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. (1995), henceforth abbreviated as BGS, hint that prosocial intentions will sometimes survive. BGS 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-