	II		
I		C	D
	C	$3,$3	$1,$4
	D	$4,$1	$2,$2

Figure 1 (Hausman). A prisoner's dilemma game form

Player II believes the same of Player I. Player I can then reason that Player II will definitely play *H,* update his or her subject probability accordingly, and play *H.* The problem lies with one idealized development of the standard view of rational belief, not with the view itself.

Many of the purported paradoxes Colman discusses yield to similar, though more complicated treatment. But some of the purported paradoxes are not paradoxes at all, and some of the apparent experimental disconfirmations are dubious. Consider first the standard single-shot prisoner's dilemma (PDG). Mutual defection is the uniquely rational outcome. Colman takes this to be paradoxical and to show that rationality is self-defeating, on the grounds that mutual cooperation is better for both players. In addition, he cites evidence showing that many experimental subjects, in fact, cooperate.

Although rationality is indeed *collectively* self-defeating in a PDG, there is no paradox or problem with the theory of rationality, and the apparently disconfirming data Colman cites are questionable. Consider the following game form (Fig. 1), which represents a PDG if the two players care only about their own monetary payoffs.

Mutual cooperators do better than mutual defectors. But the benefit comes from the choice the other player makes, not from one's own choice. (Remember this is a simultaneous play one-shot game in which I and II choose independently.) Unlike the finite iterated prisoner's dilemma or the centipede game, mutual cooperators cannot taunt mutual defectors, "If you're so rational, how come you ain't rich?" because the defectors can reply, "Because I wasn't lucky enough to be playing against a fool."

In addition, the apparently disconfirming experimental evidence is dubious, because cooperating subjects facing a game form like the one in Figure 1 might not be playing a PDG. To know what game they are playing one needs to know their preferences. For example, unless II prefers the outcome where II gets $4 and I gets $1 to the actual outcome of $3 each, II was not playing a PDG. For those who do not have these preferences, the interaction depicted in Figure 1 is not a prisoner's dilemma. Similar remarks apply to the tetrapod in Colman's Figure 5. If the numbers represented dollars, many people would prefer the outcome where both get $18 and player I's trust is rewarded, to the outcome where II gets $19 and I gets $8. The numbers in Figure 5 are, of course, supposed to represent utilities rather than dollars, but the common view, that the recommendation to play down on the first move is absurd, may reflect a common refusal to believe that these numbers correctly represent the preferences.

A great deal remains to be done to figure out how to represent rational beliefs. Wonderful controversy still rages. But one should not thereby conclude, as Colman does, that "the conception of rationality on which it [game theory] rests appears to be internally deficient" (target article, sect. 9.2). His essay does not address the treatment of rational preference, and the problems Colman explores concerning rational belief show, at most, the limitations of specific modeling choices, rather than a deficiency in basic concepts.[1]

NOTE

[1] I do not, in fact, think that the standard theory of rationality is unproblematic (see e.g., my 1992 book, Chs. 2, 12, 13), but the difficulties I see are independent of those that Colman alleges.

The limits of individualism are not the limits of rationality

Susan Hurley
PAIS, University of Warwick, Coventry CV4 7AL, United Kingdom.
susan.hurley@warwick.ac.uk **www.warwick.ac.uk/staff/S.L.Hurley**

Abstract: Individualism fixes the unit of rational agency at the individual, creating problems exemplified in Hi-Lo and Prisoner's Dilemma (PD) games. But instrumental evaluation of consequences does not require a fixed individual unit. Units of agency can overlap, and the question of which unit should operate arises. Assuming a fixed individual unit is hard to justify: It is natural, and can be rational, to act as part of a group rather than as an individual. More attention should be paid to how units of agency are formed and selected: Are the local processes local or nonlocal? Do they presuppose the ability to understand other minds?

I disagree with little that Colman says about the limitations of orthodox rational choice theory, but wonder why he doesn't say more to challenge individualism as their source, and why he omits references to trailblazers such as Regan (1980) and Howard (1988).

In 1989, I argued that Hi-Lo and Prisoner's Dilemma games (PDs) exemplify the limits of individual rationality. In Hi-Lo, individuals have the same goals, yet individual rationality fails to guarantee them the best available outcome. In PDs, individuals have different goals, and individual rationality guarantees an outcome worse for all than another available outcome. These problems stem not from nature of individuals' goals, or the instrumental character of rationality, but from individualism about rationality, which holds the unit of rational agency exogenously fixed at the individual (cf. Hurley 1989).

Activity by a given unit of agency has consequences, calculated against a background of what occurs outside that unit, and can be evaluated instrumentally. Such consequentialist evaluation does not require the unit whose activity is evaluated to be fixed at the individual. Larger units of agency can subsume smaller ones, and consequentialist evaluation can apply to different units, with different results. We can think of individuals as composed of persons-at-times (or in other ways, involving multiple personalities); similarly, we can think of collective agents as composed of persons. In both cases, lower-level rationality (or irrationality) may coexist with, or even explain, higher-level irrationality (or rationality). For example, we understand from social dilemmas and social choice theory how a group can behave irrationally as a unit, although the agents composing it are individually rational. Intrapersonal analogues of social dilemmas may explain some forms of individual irrationality. Conversely, agents can behave irrationally as individuals, yet their actions fit together so that the group they compose behaves rationally (Hutchins 1995, pp. 235ff).

Individualism requires the individual to do the individual act available that will have the best expected consequences, given what other individuals are expected to do. Given others' expected acts, an individual agent has certain possible outcomes within her causal power. The best of these may not be very good, and it may be indeterminate what others are expected to do. But a group of individuals acting as a collective agent can have different possible outcomes within its causal power, given what agents outside the group are expected to do. A collective agent may be able to bring about an outcome better than any that the individual agent can bring about – better for that individual, inter alia. If so, the issue is not just what a particular unit of agency should do, given others'

https://doi.org/10.1017/S0140525X03340051 Published online by Cambridge University Press

expected acts, but also *which* unit should operate. The theory of rationality has yet to endogenize the latter question; Bacharach calls this "an important lacuna" (1999, p. 144; but cf. Regan 1980).

The assumption of a fixed individual unit, once explicitly scrutinized, is hard to justify. There is no theoretical need to identify the unit of agency with the source of evaluations of outcomes; collective agency does not require collective preferences. Although formulations of team reasoning may assume team preferences (see target article, sect. 8.1), what is distinctive about collective agency comes into sharper relief when it is made clear that the source of evaluations need not match the unit of agency. As an individual, I can recognize that a wholly distinct agent can produce results I prefer to any I could bring about, and that my own acts would interfere. Similarly, as an individual I can recognize that a collective agent, of which I am merely a part, can produce results I prefer to any I could bring about by acting as an individual, and that my doing the latter would interfere. Acting instead in a way that partly constitutes the valuable collective action can be rational. Not only can it best serve my goals to tie myself to the mast of an extended agent, but rationality itself can directly so bind me – rather than just prompt me to use rope.

Acting as part of a group, rather than as an individual, can also be natural. Nature does not dictate the individual unit of agency. Persons can and often do participate in different units, and so face the question of which unit they *should* participate in. Moreover, the possibility of collective agency has explanatory power. For example, it explains why some cases (e.g., Newcomb's Problem and Quattrone & Tversky's voting result) of supposedly evidential reasoning have intuitive appeal, while others (e.g., the smoking gene case) have none (Hurley 1989, Ch. 4; 1991; 1994).[1]

If units of agency are not exogenously fixed, how are units formed and selected? Is centralized information or control required, or can units emerge as needed from local interactions? At what points are unit formation and selection rationally assessable? I cannot here offer a general view of these matters, but highlight two important issues.

First, are the relevant processes local or nonlocal? Regan's version of collective action requires cooperators to identify the class of those intending to cooperate with whomever else is cooperating, to determine what collective action by that group would have the best consequences (given noncooperators' expected acts), and then play their part in that collective action. This procedure is nonlocal, in that cooperators must type-check the whole class of potential cooperators and identify the class of cooperators before determining which act by that group would have the best consequences. This extensive procedure could be prohibitive without central coordination. The problem diminishes if cooperators' identities are preestablished for certain purposes, say, by their facing a common problem, so preformed groups are ready for action (see Bacharach 1999).

A different approach would be to seek local procedures from which potent collective units emerge. Flexible self-organization can result from local applications of simple rules, without central coordination. Slime mold, for example, spends most of its life as separate single-celled units, but under the right conditions these cells coalesce into a single larger organism; slime mold opportunistically oscillates between one unit and many units. No headquarters or global view coordinates this process; rather, each cell follows simple local rules about the release and tracking of pheromone trails.

Howard's (1988) Mirror Strategy for one-off PDs may allow groups of cooperators to emerge by following a simple self-referential local rule: Cooperate with any others you encounter who act on this very same rule. If every agent cooperates just with its copies, there may be no need to identify the whole group; it may emerge from decentralized encounters governed by simple rules. Evidently, rules of cooperation that permit groups to self-organize locally have significant pragmatic advantages.

Both Regan's and Howard's cooperators need to perceive the way one another thinks, their methods of choice. Which choices their cooperators make, depends on which other agents are cooperators, so cooperation must be conditioned on the *methods* of choice, not the choices, of others. If method-use isn't perfectly reliable, however, cooperators may need to be circumspect in assessing others' methods and allow for the possibility of lapses (Bacharach 1999).

These observations lead to the second issue I want to highlight: What is the relationship between the processes by which collective agents are formed and selected, and the ability to understand other minds? Does being able to identify with others as part of a unit of agency, require being able to identify with others mentally? Psychologists ask: What's the functional difference between genuine mind-reading and smart behavior-reading (Whiten 1996)? Many social problems that animals face can be solved merely in terms of behavior-circumstance correlations and corresponding behavioral predictions, without postulating mediating mental states (see Call & Tomasello 1999; Heyes & Dickinson 1993; Hurley 2003; Povinelli 1996). What kinds of problems also require understanding the mental states of others?

Consider the kinds of problems that demonstrate the limitations of individualistic game theory. When rational individuals face one another, mutual behavior prediction can break down in the ways that Colman surveys; problem-solving arguably requires being able to understand and identify with others mentally. If cooperators need to know whether others have the mental processes of a cooperator before they can determine what cooperators will do, they must rely on more than unmediated associations between circumstances and behavior. Collective action would require mind-reading, not just smart behavior-reading. Participants would have to be mind-readers, and be able to identify, more or less reliably, other mind-readers.

NOTE

[1] It is widely recognized that Prisoners' Dilemma can be interpreted evidentially, but less widely recognized that Newcomb's Problem and some (but not all) other cases of supposed evidential reasoning can be interpreted in terms of collective action.

Coordination and cooperation

Maarten C. W. Janssen

Department of Economics, Erasmus University, 3000 DR, Rotterdam, The Netherlands. **janssen@few.eur.nl** **www.eur.nl/few/people/janssen**

Abstract: This comment makes four related points. First, explaining coordination is different from explaining cooperation. Second, solving the coordination problem is more important for the *theory* of games than solving the cooperation problem. Third, a version of the Principle of Coordination can be rationalized on individualistic grounds. Finally, psychological game theory should consider how players perceive their gaming situation.

Individuals are, generally, able to get higher payoffs than mainstream game-theoretic predictions would allow them to get. In coordination games, individuals are able to coordinate their actions (see e.g., Mehta et al. 1994a; 1994b; Schelling 1960) even though there are two or more strict Nash equilibria. In Prisoner's Dilemma games, individuals cooperate quite often, even though mainstream game theory tells that players should defect. In this comment, I want to make four points. First, it is important to distinguish the cooperation problem from the coordination problem. Second, from the point of view of developing a *theory* of games, the failure to explain coordination is more serious than the failure to explain cooperation. Third, the Principle of Coordination, used to explain why players coordinate, can be rationalized on individualistic grounds. One does not need to adhere to "we thinking" or "Stackelberg reasoning." Finally, psychological game theory may gain predictive power if it takes into account how players perceive their gaming situation.

https://doi.org/10.1017/S0140525X03340051 Published online by Cambridge University Press