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.

The problem of *coordination* is different from the problem of *cooperation.* In a cooperation problem, as the *one-shot* Prisoner's Dilemma, players have a dominant strategy, which is *not* to cooperate, and one may wonder why people deviate from their dominant strategy and *do* cooperate. To explain cooperation, one has to depart from the axiom of individual rationality. This is not the case for the problem of coordination. In a coordination problem, there are two or more Nash equilibria in pure strategies and the issue is that individual rationality considerations are *not sufficient* to predict players' behavior. To explain coordination, an approach that supplements the traditional axioms of individual rationality may be taken.

In a truly *one-shot* Prisoner's Dilemma, where the payoffs are formulated such that players care only about their individual payoffs, I find it hard to find reasons (read: to explain) why people cooperate. Of course, I don't want to deny the empirical evidence, but the dominant strategy argument seems to me very appealing and difficult to counteract. If people choose to cooperate, they must be in one way or the other boundedly rational. I think the *theory* of games should not just explain how people in reality behave when they play games. It should also have an answer to the question *why,* given their own preferences, they behave in a certain way. The weakest form this requirement can take is that, given a theoretical prediction people understand, it is in their own interest not to deviate from the prediction. In other words, a theory of games should be reflexive. The problem with a theory of games which says that players cooperate is that "smart students" don't see any reason why it is beneficial to do so. If the theory makes a prediction, then it should be in the interest of the players to make that prediction come true.

In coordination problems, the concept of Nash equilibrium is too weak, as it does not give players a reason to choose one out of several alternatives. Gauthier (1975), Bacharach (1993), Sugden (1995), and Janssen (2001b) make use of (a version of) the Principle of Coordination to explain coordination. Janssen (2001a) develops a relatively simple framework that rationalizes the uniqueness version of this Principle. The basic idea is that each player individually forms a plan, specifying for each player how to play the game, and which conjecture to hold about their opponent's play. Individual plans should satisfy two axioms. *Individual rationality* says that a plan be such that the sets of strategies that are motivated by the plan must be best responses to the conjectures that are held about the other player's play. *Optimality* requires that players formulate optimal plans, where a plan is optimal if the maximum payoff both players get if they follow this plan is larger than the minimum payoff both players would get according to any alternative plan satisfying the individual rationality axiom.

If there is a unique strict Pareto-efficient outcome, then there is a unique plan satisfying Individual Rationality and Optimality how to play the game. To see the argument, consider the following game (Table 1).

It is clear that a plan where every player conjectures the other to play *L,* and where both players actually choose *L,* is a plan that satisfies Individual Rationality and, moreover, is better for both players than any other plan. As the plan is uniquely optimal, both players *thinking individually* formulate the same plan, and they will choose to do their part of it.

Note that the above approach is different from "we thinking" as discussed by Colman, as no common preferences are specified.

Table 1 (Janssen). *A game of pure coordination with a uniquely efficient equilibrium*

	L	*R*
L	*2,2*	*0,0*
R	*0,0*	*1,1*

Table 2 (Janssen). *A game of pure coordination without a uniquely efficient equilibrium*

	Blue	*Blue*	*Red*
Blue	*1,1*	*0,0*	*0,0*
Blue	*0,0*	*1,1*	*0,0*
Red	*0,0*	*0,0*	*1,1*

Also, no coach is introduced who can make recommendations to the players about how to coordinate their play, as in Sugden (2000, p. 183).

This approach, by itself, cannot explain coordination in a game where two players have to choose one out of three (for example, two blue and one red) objects and where they get awarded a dollar if they happen to choose the same object. Traditionally, game theory would represent this game in the following "descriptively objective" matrix (Table 2).

Intuitively, the players should pick the red object, but the Principle of Coordination advocated here, by itself, cannot explain this intuition.

Psychological game theory may, in addition to the elements mentioned by Colman, also further investigate Bacharach's (1993) suggestion, and investigate how people describe the game situation to themselves (instead of relying on some "objective" game description). By using the labels of the strategies, individuals may describe the above game as being a game between picking a blue and a red object, where the chance of picking the *same* blue object, given that both pick a blue object, is equal to a half. Given such a description, there is (again) a unique plan satisfying Individual Rationality and Optimality.

Which is to blame: Instrumental rationality, or common knowledge?

Matt Jones and Jun Zhang

Department of Psychology, University of Michigan, Ann Arbor, MI 48109-1109. **mattj@umich.edu** **junz@umich.edu**
http://umich.edu/~mattj

Abstract: Normative analysis in game-theoretic situations requires assumptions regarding players' expectations about their opponents. Although the assumptions entailed by the principle of common knowledge are often violated, available empirical evidence – including focal point selection and violations of backward induction – may still be explained by instrumentally rational agents operating under certain mental models of their opponents.

The most important challenge in any normative approach to human behavior is to correctly characterize the task the person is presented with. As Colman points out, the normative analysis of game settings provided by instrumental rationality is incomplete; information must be included about the opponent. We argue here that the common knowledge of rationality (CKR) axioms, which are meant to extend normative analysis to game theory, actually limit the rationality attributed to subjects. When players are allowed to reason about their opponents, using more information than just that provided by CKR2, we find that the major phenomena cited as evidence against rational choice theory (RCT) – focal point selection and violations of backward induction arguments – can be predicted by the resulting normative theory. This line of reasoning follows previous research in which supposed suboptimalities in human cognition have been shown to be adaptive given a more fully correct normative analysis (e.g., Anderson &

https://doi.org/10.1017/S0140525X03350058 Published online by Cambridge University Press