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 &

Schooler 1991; Anderson et al. 1997; Flood 1954; Jones & Sieck, in press; Oaksford & Chater 1996; Schacter 1999).

The difficulty with the CKR axioms is that they require players to reason about their opponents entirely a priori, based only on the assumptions of rationality and common knowledge, while ignoring all other potential sources of information. A more faithful model of rational choice would allow players to utilize all the knowledge available to them, including general knowledge about human behavior or specific knowledge about the opponent gained from previous interactions (e.g., earlier moves). For example, the fact that the conventional priority of Heads over Tails leads to the phenomenon of focal point selection should realistically be available to each player as information for use in predicting the opponent's choice. Thus, all that is needed is a simple intuitive understanding of human behavior for a subject to infer correctly (and rationally) that the opponent is likely to choose the focal option. Instrumental rationality then dictates that the player chooses that option as well. Similar reasoning applies to payoff dominance in the case of the Hi-Lo matching game.

Relaxing the restrictions provided by CKR on players' models of their opponents can also explain violations of the prescriptions of backward induction arguments. If Player II's model of Player I admits alternatives to perfect rationality, then an initial cooperative move by Player I will simply lead to an update of II's beliefs about I (rather than generating a logical impasse). This sort of updating can be formalized using a Bayesian framework, in which each player has probabilistic prior beliefs about the opponent (perhaps peaked around rationality, but nonzero elsewhere), which are determined by prior experience with the opponent or with people in general. Even if the prior expectation were heavily biased towards strict rationality, an initial cooperative move by Player I would force Player II's model to favor other possibilities, for example, that Player I always plays Tit-For-Tat. This could lead to Player II cooperating on step 2, in turn giving Player I justification for cooperating on step 1.

The preceding arguments have shown how failures of CKR can be remedied by more complete normative analyses that preserve the assumption of instrumental rationality, that is, optimality of actions as conditioned on the model of the opponent. The question of rationality in game scenarios then shifts to the rationality of that model itself (inductive rationality). In the case of focal point selection, we have offered no specific mechanism for the inductive inference regarding the opponent's likely choice, as based on general experience with human behavior. We merely point out that it is perfectly consistent with the assumption of inductive rationality (although it has no basis in CKR). (Ironically, the same empirical fact that is cited as evidence against RCT – namely, focal point selection – actually corroborates the rationality of people's inductive inferences.)

The stance taken in our discussion of backward induction, whereby people are rational yet they entertain the possibility that others are not, presents a subtler problem. What must be remembered here is that, as a positive theory, RCT only claims that people try to act rationally (target article, sect. 3.3), and that the idealization of perfect rationality should give qualitatively correct predictions. Of course, in reality, people do err, and subjects are aware of this fact. Therefore, in forming expectations about their opponents' actions, subjects are open to the possibility of errors of reasoning by the opponent. Furthermore, as one progresses further back in the chain of reasoning entailed by backward induction, the expectation of such errors compounds. Thus, the framework proposed here can be viewed as idealizing rationality at the zero level, but not at higher orders of theory-of-mind reasoning.

Our thesis, that people follow instrumental rationality but anchor it on their model of the opponent, is supported by Hedden and Zhang's (2002) recent investigation of the order of theory-of-mind reasoning employed by subjects in three-step sequential-move games. On each trial, subjects, who controlled the first and third moves, were asked first to predict the response of the opponent (a confederate who controlled the second move) and their own best choice on the first move. Initially, subjects tended to predict myopic choices by the opponent, corresponding to level 0 reasoning (level 1 was optimal for the opponent). Accordingly, subjects' own actions corresponded to the level 1 strategy, rather than the level 2 strategy prescribed by CKR. However, after sufficient experience with an opponent who played optimally, 43% of subjects came to consistently predict the opponent's action correctly, and altered their own behavior to the level 2 strategy. Although the remaining subjects failed to completely update their mental model of the opponent, errors of instrumental rationality (discrepancies between the action chosen and that dictated by the expectation of the opponent's response) remained low and approximately constant throughout the experiment for both groups. These results support the claim that violations of the predictions of CKR can be explained through scrutiny of player's models of their opponents, without rejecting instrumental rationality, and suggest that further investigations of rational choice in game situations must take into account the distinction between instrumental and inductive rationality.

Analogy in decision-making, social interaction, and emergent rationality

Boicho Kokinov

Central and East European Center for Cognitive Science, Department of Cognitive Science and Psychology, New Bulgarian University, Sofia, 1618 Bulgaria. **bkokinov@nbu.bg**
http://www.nbu.bg/cogs/personal/kokinov

Abstract: Colman's reformulation of rational theory is challenged in two ways. Analogy-making is suggested as a possible candidate for an underlying and unifying cognitive mechanism of decision-making, one which can explain some of the paradoxes of rationality. A broader framework is proposed in which rationality is considered as an emerging property of analogy-based behavior.

Rationality has long been shown to fail as a descriptive theory of human decision-making, both at the individual and social levels. In addition, Colman presents strong arguments that rationality also fails as a normative theory for "good" decision-making – "rational" thinking does not produce optimal behavior in social interaction and even acts against the interests of the individual in some cases. Fortunately, human beings often act against the postulates of rationality and achieve better results than prescribed by the theory. Therefore, Colman concludes that "rationality" has to be redefined by extending it with additional criteria for optimization, such as the requirement for maximizing the "collective" payoff, or with additional beliefs about the expected strategies of the coplayers. He does not clarify how and when these additional criteria are triggered or where the common beliefs come from.

We are so much attached to the notion of rationality that we are always ready to repair it, but not to abandon it. The theory of rationality is, in fact, a formalization of a naive theory of human thinking. This naive theory makes it possible to predict human behavior in most everyday situations in the same way as naive physics makes it possible to predict natural phenomena in everyday life. However, no one takes naive physics so seriously as to claim that it provides "the explanation" of the world. Moreover, even refined and formalized versions of this naive theory, like Newtonian mechanics, are shown not to be valid; and more complicated and counterintuitive theories at the microlevel, like quantum mechanics, have been invented. On the contrary, rationality theory is taken seriously, especially in economics, as an explanation of human behavior.

Instead of extending rationality theory with additional socially oriented rules, it may be more useful to make an attempt to build a multilevel theory that will reveal the implicit and explicit cognitive processes involved in decision-making. These underlying cog-