Paradoxes of Interactive Rationality: A Unified View
Extended Abstract

Eric Pacuit* Olivier Roy†

An increasingly popular, but of course not uncontroversial[1], view is that "*the* fundamental insight of game theory [is] that a rational player must take into account that the players reason about each other in deciding how to play" (Aumann and Dreze, 2008, pg. 81). Exactly how the players (should) incorporate the fact that they are interacting with other (actively reasoning) agents into their own decision making process is the subject of much debate. A variety of frameworks have been put forward to explicitly model the *reasoning* of rational agents in a strategic situation. Key examples include Brian Skyrms' models of "dynamic deliberation" (Skyrms, 1990), Ken Binmore's analysis of "eductive reasoning" (Binmore, 1987), and Robin Cubitt and Robert Sugden's "common modes of reasoning" (Cubitt and Sugden, 2011).Although the details of these frameworks are quite different[2], they share a common line of thought: Contrary to classical game theory, *solution concepts* are no longer the basic object of study. Instead, the "rational solutions" of a game are the result of individual (rational) decisions in specific informational "contexts".

This perspective on the foundations of game theory is best exemplified by the so-called epistemic program in game theory (cf. Brandenburger, 2007). The central thesis here is that the basic mathematical model of a game should include an explicit parameter describing the players' *informational attitudes*. However, this broadly decision-theoretic stance does not simply *reduce* the question of decision-making in interaction to that of rational decision making in the face of uncertainty. Crucially, *higher-order* information (belief about beliefs, etc.) are key components of the informational context of a game:

> "In any particular structure, certain beliefs, beliefs about belief, ..., will be present and others won't be. So, there is an important implicit assumption behind the choice of a structure. This is that it is "transparent" to the players that the beliefs in the type structure — and only those beliefs — are possible ....The idea is that there is a "context" to the strategic situation (eg., history, conventions, etc.) and this "context" causes the players to rule out certain beliefs."
>
> (Brandenburger and Friedenberg, 2010, pg. 801)

Of course, different contexts of a game can lead to drastically different outcomes. But this means that the informational contexts themselves are open to rational criticism:

> "It is important to understand that we have two forms of irrationality in this paper...For us, a player is rational if he optimizes and also rules nothing out. So irrationality might mean not optimizing. But it can also mean optimizing while not considering everything possible."
>
> (Brandenburger et al., 2008, pg. 314)

*Tilburg Institute for Logic and Philosophy of Science, Tilburg University, e.j.pacuit@uvt.nl

†Center for Mathematical Philosophy, LMU, Olivier.Roy@lrz.uni-muenchen.de

[1]Consider, for example, the following quote from Kadane and Larkey (1982): "It is true that a subjectivist Bayesian will have an opinion not only on his opponent's behavior, but also on his opponent's belief about his own behavior, his opponent's belief about his belief about his opponent's behavior, etc. (He also has opinions about the phase of the moon, tomorrow's weather and the winner of the next Superbowl.) However, in a single-play game, all aspects of his opinion except his opinion about his opponent's behavior are irrelevant, and can be ignored in the analysis by integrating them out of the joint opinion." (pg, 239).

[2]Skyrms assumes the players deliberate by calculating their *expected utility* and then use this new information to recalculate their probabilities about the states of the world and recalculate their expected utilities. Binmore models the players as *Turing machines* that can *compute* their rational choices. And, Cubitt and Sugden build on David Lewis' analysis of common knowledge in terms (inductive/deductive) rules that are commonly accepted among all the players.

So, a player can be rationally criticized for not choosing what is *best* or what one *ought to* choose, *given one's information*. What counts as "best" or what ought to be done will be determined by the *reasons* that the players have and by the *normative facts* that hold *in* a given context.[3] But, a player can also be rationally criticized for not reasoning *to* a "proper" context. What counts as a "proper" context is still the subject of much debate. The point is that there is rational pressure for a player to "properly ignore" states where at least one of the players does not choose optimally without making *substantive assumptions* about the beliefs of her opponents.[4]

In this paper, we are interested in this second "form" of rationality (eg., how does a rational player reason *to* a "proper" context?). Modern work in dynamic logics of belief revision provides us with a rich repertoire of notions of belief (eg., safe belief, strong belief) and informative actions (eg., radical upgrade, conservative upgrade).[5] Building on recent results about long-term dynamics of beliefs (Baltag and Smets, 2009), we will employ these dynamic logics of belief revision to provide a fresh look at key issues in the epistemic foundations of game theory. In particular, we are interested in issues surrounding the epistemic characterization of *iterated elimination of weakly dominated strategies* (IEWDS). The goal is not to provide an alternative epistemic characterization of this solution concept (both Brandenburger et al. (2008) and Halpern and Pass (2009) have convincing results here), but rather to explore the dynamics of the epistemic models that lead to such characterization results.

We only have the space here to explain the main idea behind our analysis. Larry Samuelson (1992) was the first to be explicit about the puzzle surrounding the epistemic foundations of IEWDS. He showed (among other things) that there is no epistemic model of the following game with at least one state satisfying "common knowledge of admissibility" (i.e., a state where there is common knowledge that the players do not play a strategy that is weakly dominated).

		Bob	
		L	R
Ann	u	$1,1$	$1,0$
	d	$1,0$	$0,1$

We will use dynamic modal logics of belief revision to offer a new perspective on this result. Initially, assume that all outcomes of the game are equally plausible. In this informational context, the assumed choice rule (in this case, "do not play non-admissible strategies") fixes the normative facts: for example, "Ann ought not to play d". This normative fact licenses an upgrade with the proposition "Ann does not play d" resulting in a model where it is commonly believed that "Ann does not play d". But this is a new informational context with new rationality requirements. In particular, "Bob ought not to play R" licenses the upgrade with the proposition "Bob does not play R" resulting in a model where it is commonly believed that "Ann plays u and Bob plays L". In this informational context (which is essentially the IEWDS solution), both u and d are *permitted* for Ann. In this case, the rational response is for the players to *suspend their belief* that Ann will not play d. Continuing in this manner, we will prove that this process of repeatedly upgrading with propositions licensed by the relevant normative facts never reaches a fixed-point, and use this to explain Samuelson's result that there is no epistemic model of the above game with common knowledge of admissibility. The main technical part of the paper will study what happens as we vary the choice rule and types of upgrades. For example, we can show that a fixed-point is always reached if the players use the choice rule "do not play a strategy that is strictly dominated".

Broader project: interactive rationality This paper is part of a broader project (Pacuit and Roy, 2011a) which looks at results in the epistemic foundations of game theory from the perspective of recent

[3]We assume that *choice rules* fix what count as reasons for or against actions, and responding correctly to these reasons will a be necessary condition for interactive rationality. There is much more to say here, but this is the subject of a companion paper (Pacuit and Roy, 2011b).

[4]This issue is explored in much more detail in (Roy and Pacuit, 2010).

[5]The reader not familiar with this area can consult the recent textbook (van Benthem, 2010) for details.

work in meta-ethics about normative requirements using the tools of dynamic epistemic logic. Our goal is to develop of theory of *interactive rationality*. Interactive rationality is not only a matter of responding correctly to reasons, it also encompasses *requirements* (Broome, 2010) of epistemic consistency and practical coherence. For single-agent situations, such requirements have been extensively studied in the last years. Social interactions, however, do raise requirements of their own. Take, for instance, the requirements of consistency between an agent's first and higher-order information, i.e., the notion of "consistent belief hierarchies" (Brandenburger and Dekel, 1993). One can also think of dynamic consistency in extensive games in terms of interactive, practical requirements (think also of commitments and credible threats in sequential plays). We take correct response to reasons *and* conformity to rational requirements to be two necessary conditions for the broader notion of interactive rationality. If an agent does not respond correctly to the reasons induced by a given choice rule we will say that she is not rational. The reader should keep in mind, however, that in general the converse does not hold: an agent can be irrational because, say, her first and higher-order beliefs are mutually inconsistent, even though she responds correctly to her reasons.

References

Aumann, R. and J. Dreze (2008). Rational expectations in games. *American Economic Review 98*, 72 – 86.

Baltag, A. and S. Smets (2009). Group belief dynamics under iterated revision: Fixed points and cycles of joint upgrades. In *Proceedings of Theoretical Aspects of Rationality and Knowledge*.

Binmore, K. (1987). Modeling rational players: Part I. *Economics and Philosophy 3*, 179 – 214.

Brandenburger, A. (2007). The power of paradox: some recent developments in interactive epistemology. *International Journal of Game Theory 35*, 465–492.

Brandenburger, A. and E. Dekel (1993). Hierarchies of beliefs and common knowledge. *Journal of Economic Theory 59*, 189–198.

Brandenburger, A. and A. Friedenberg (2010). Self-admissible sets. *Journal of Economic Theory 145*, 785 – 811.

Brandenburger, A., A. Friedenberg, and H. J. Keisler (2008). Admissibility in games. *Econometrica 76*, 307–352.

Broome, J. (2010). Rationality through reasoning. Manuscript.

Cubitt, R. and R. Sugden (2011). Common reasoning in games: A Lewisian analysis of common knowledge of rationality. CeDEx Discussion Paper.

Halpern, J. and R. Pass (2009). A logical characterization of iterated admissibility. In A. Heifetz (Ed.), *Proceedings of the Twelfth Conference on Theoretical Aspects of Rationality and Knoweldge*, pp. 146 – 155.

Kadane, J. B. and P. D. Larkey (1982). Subjective probability and the theory of games. *Management Science 28*(2), 113–120.

Pacuit, E. and O. Roy (2011a). Interactive rationality. Unpublished Manuscript.

Pacuit, E. and O. Roy (2011b). Interactive rationality and the dynamics of reasons. Unpublished Manuscript.

Roy, O. and E. Pacuit (2010). Substantive assumptions and the existence of universal knowledge structures: A logical perspective. Under submission.

Samuelson, L. (1992). Dominated strategies and common knowledge. *Game and Economic Behavior 4*, 284 – 313.

Skyrms, B. (1990). *The Dynamics of Raitonal Deliberation*. Harvard University Press.

van Benthem, J. (2010). *Logical Dynamics of Information and Interaction*. Cambridge University Press.