Jean-Jacques Laffont Lecture
2010

Stringent Criteria for Rational Strategic Behavior

Robert Wilson

Prolog

I am glad to participate in this annual occasion to remember Jean-Jacques Laffont. His published works are permanent contributions to the heritage of economics. His central role in building IDEI and the Toulouse School of Economics into an institution with international prominence brings lasting benefits. For those of us who knew him personally, his lively intellect and qualities of character, integrity, and compassion will always be remembered fondly. It is very sad that his vigorous and productive life ended abruptly when he was at the height of his intellectual power.

My last memory of him was a pleasant occasion here in Toulouse when he (with Jacques Crémer, Jean Tirole, and I) donned medieval academic regalia for the defense of the doctoral dissertation by his student Juan Carrillo – I will always remember that vivid image of Jean-Jacques as a teacher, scholar, and the very soul of the academic institution he founded.

Introduction

My lecture today reports on recent developments in a topic that interested Jean-Jacques in his work on incentives during the 1980s. That subject is – what exactly characterizes fully rational behavior in strategic interactions? That is, can there be a theory of rational decision making in multi-person contexts that is nearly as complete as it is for single-person contexts, notably in the work of John von Neumann (1943) and Leonard J. Savage (1954). Although important contributions were made by various authors, this difficult question was never fully answered in the 1970s and 1980s, and subsequently it largely disappeared from the research agenda of economic theory. Yet it remains central to the foundations of economics.

The question is posed precisely in a seminal article by Elon Kohlberg and Jean-François Mertens (1986). Their article responds to the many proposed refinements of John Nash's (1951) definition of equilibrium in a well-defined game played by multiple participants. These proposals address some deficiencies of Nash's definition, but each proposal as its own deficiencies. Kohlberg and Mertens argue that the question will be answered only by showing that some simple basic axioms characterize exactly what the right definition should be. Indeed, it was the axiomatic method that characterized rationality of single-person decisions.

The question is then as much a matter of philosophy as it is of economics or the other social sciences. For, this program demands that we isolate some elementary properties that we believe must be true for rational strategic behavior, and then requires further that we show that these elementary properties imply the other properties that one attributes to rational strategic behavior.

But if it is a matter of philosophy, and even logic, is this exercise useful for economics? That the answer is yes can be seen in the fact that most economic theory (and study by a student of economics) is cast presently in terms of models formulated as games. Game theory provides a precise language for modeling multi-person interactions and a precise method for analyzing them, deriving predictions, and testing them with experimental and empirical data.

As at virtually all major institutions for economic research, at the Toulouse School of Economics researchers and students routinely apply game theory in their work on designs of contracts, markets, and regulations. So, it is important to establish the foundations of game theory, even if we are late to the task considering that applications of game theory have forged ahead without firm foundations.

Synopsis

In today's lecture I consider only finite dynamic games with just two players, and that satisfy two technical conditions: perfect recall (a player never forgets what he or she knew previously), and generic payoffs (players' payoffs are not linked by algebraic equations).

I describe four axioms that exactly characterize a very stringent definition of rational behavior in strategic interactions. This definition turns out to be Mertens' (1991) definition of stability. Actually, stability is the most stringent criterion among the dozens that have been proposed.

The results I present were obtained in work with Srihari Govindan in a collaboration that has now extended over fifteen years. I feel enormous gratitude to Srihari for enabling me to participate in this project.

Background

Game theory has been criticized vigorously. Here I argue briefly that much has been learned from the critics.

One criticism is that in life one rarely encounters a situation posed precisely as a game – in particular, even if there is a game, it is not 'common knowledge' among the participants. Those who apply game theory have learned from this criticism to adopt more general models, for example, models with incomplete information about preferences and available strategies – even models with unawareness about others' preferences and strategies. One strand of game theory formulates 'epistemic' models that represent precisely the extent of common knowledge among participants and its implications.

A second criticism is that game theory assumes too much rationality – it is evident that human rationality is inherently limited by bounded memory and bounded processing capability, and also affected by emotional and contextual factors, as well as societal influences such as expectations of fairness. In fact, the evidence from experiments is that game-theoretic models are imperfect predictors. Practitioners of game theory do indeed yearn to study models with bounded memory and processing, but there have been few applications because they are more difficult to formulate and analyze. Similarly, one can argue that, in principle, emotions and fairness are aspects of preferences that can be modeled; for instance, there has been substantial success in modeling habit formation and dependence of decisions on the status quo, such as loss aversion and ambiguity aversion.

So, regarding these criticisms, I conclude that their impact on game theory has been to encourage use of better models of memory and processing, and richer models of preferences.

But there is one criticism that is not often addressed adequately. This is the observation that one rarely plays a game in isolation. In life, every strategic interaction is embedded in a wider context. Besides the players in a narrowly circumscribed game with their strategies, outcomes, and preferences, there are usually other participants who have strategies, outcomes, and preferences – it is just that these other players are peripheral to the particular game considered. Savage (1954) and especially Mertens (1992) describe such a situation by calling the game a ‘small world’ embedded in a larger context, which here I will call a ‘metagame’.

In cognitive and social psychology, Daniel Kahneman and Amos Tversky (2000) and Lee Ross (1991) are prominent proponents of the relevance of the significant effects of embedding interactions in wider contexts. In their terminology, embedding a game in a larger metagame creates ‘framing’ or ‘presentation’ effects. Framing can occur at the level of an individual, such as how a decision maker perceives the status quo when assessing gains and losses. In multi-person interactions, framing can affect which decisions are chosen, such as when the signaling implications of others’ actions are interpreted in the richer context of the metagame.

So, one might conclude that due to framing effects it is not possible to analyze any one game without considering also the metagame in which it is embedded, the meta-metagame in which the metagame is embedded, and so on until one must deal with a model of everything.

But this is not my conclusion. If the other players in metagames are truly peripheral, then the decisions of a fully rational player in the original game should not be affected. This is the motivation for a key axiom that I describe later. Essentially, the axiom excludes framing effects. Of course one must still deal with the evidence from experiments that players are vulnerable to framing effects – but here our aim is solely to describe the behavior of fully rational players who are immune to framing effects.

Axioms

I now describe the four axioms used to characterize rational behavior.

0. **Equilibrium.** A player's strategy is an optimal reply to the other's strategy.

Axiom 0 is just John Nash's (1951) definition of an equilibrium. It has the well-known deficiency that it allows a player to use a strategy that is dominated. That is, some other strategy always does as well, and in some cases better. So the next axiom excludes use of dominated strategies.

1. **Undominated Strategies.** No player uses a dominated strategy.

The next axiom assumes that the prediction might take the form of a connected closed set of equilibria. Kohlberg and Mertens show that an axiomatic development must allow set-valued predictions, but here it is innocuous because we assume the game has perfect recall and payoffs are generic. Kreps and Wilson (1982) show that in this case all equilibria in a connected set have the same outcome; that is, these equilibria have the same behavior along paths of equilibrium play and differ only in responses to deviant behaviors.

A second deficiency of Nash's definition is that in a game with perfect information an equilibrium can differ from the one constructed by backward induction. More generally, rational behavior should satisfy sequential rationality –from any point in the game the continuation of a player's strategy should be optimal given some consistent belief about what the other's strategy might be in the continuation. Since the prediction is a set of equilibria with the same outcome, it suffices to require only that some equilibrium in the set satisfies sequential rationality.

Here we impose a more stringent criterion for sequential rationality than invoked in Kreps and Wilson's definition of sequential equilibrium.[1] Although Eric van Damme (1982) gave it the awkward name 'quasi-perfect' equilibrium, it is an eminently sensible strengthening of sequential rationality.

In such an equilibrium, each player's behavior is described initially by a finite sequence of strategies, interpreted as the other player's alternative hypotheses about his behavior, ordered lexicographically. But whenever a deviation from equilibrium play is detected, hypotheses that fail to explain the deviation are discarded. In each event during the game, a player continues with the first strategy in his sequence that does not exclude that event, and this continuation must be optimal in reply to the subsequence consisting of those hypotheses about the other player that do not exclude the event. Moreover, optimality here means that ties are resolved lexicographically; viz., any alternative strategy that is superior in reply to one hypothesis is inferior in reply to some hypothesis earlier in the subsequence.

2. **Backward Induction.** A prediction includes a quasi-perfect equilibrium.

A quasi-perfect equilibrium does not use dominated strategies, so it is consistent with Axiom 1. When payoffs are generic, as assumed here, each sequential equilibrium is quasi-perfect.

The last axiom excludes framing effects. First we define a metagame. Recall that it is a larger game with additional players, provided the original players' strategies and payoffs are preserved. So we require that, regardless of what the additional players' strategies are:

a. In the metagame the original players have copies of their original pure strategies that preserve their original payoffs, and

b. All their other strategies in the metagame are redundant in that each is equivalent to some mixed strategy in the original game.

[1] Govindan and I have work in progress that enables Axiom 2 to require only a sequential equilibrium, provided a stronger version of Axiom 3 is used. However, for games with generic payoffs, as assumed here, every sequential equilibrium is quasi-perfect as required by Axiom 2.

The salience of second part stems from the possibility that an original player conditions his strategy on what additional players choose, but embedding requires that this conditioning only substitutes for randomization.

3. **Exclusion of Framing Effects.** A prediction is immune to embedding in a metagame.

Note that I say "A prediction" since there is no presumption that there will be a unique prediction. It is intrinsic to game theory, and social choice theory generally, that one cannot predict a unique outcome of every game based solely on players' preferences.

Theorem

I now state the main implication of the axioms. For this I invoke Mertens' (1991) definition of a stable prediction, which is the most stringent refinement of Nash equilibrium ever proposed. His definition is very technical and phrased in terms of homology theory.[2] However, you can think of it as a procedure for computing equilibria that satisfy the axioms; in particular, a stable prediction is one for which the topological degree of a certain projection map is not zero.

For a concise statement I suppose that two strategies are identified if they exclude the same outcomes.

Theorem: The axioms are equivalent to requiring a prediction to be stable.

To interpret this theorem and show its significance, I describe the further implications of stability. These include all the properties that Kohlberg and Mertens argued should be satisfied by an axiomatic theory.

a. [Perfection] Every nearby game obtained by perturbing players' strategies has a nearby equilibrium. This continuity property assures that predictions are not vulnerable to small variations in behavior. It is the source of the name 'stable'.

[2] A simplified rendition is that a connected closed set of equilibria is stable if the projection map, from a connected closed neighborhood in the graph of equilibria over the space of players' strategies perturbed toward mixed strategies, is essential, viz., has a point of coincidence with every continuous map having the same domain and range.

b. [Invariance and Properness] A stable prediction depends only on the ‘reduced normal form’ of the game obtained by eliminating pure strategies that are redundant because they are payoff-equivalent to mixtures of other pure strategies. Moreover, a stable prediction contains a proper equilibrium (Myerson, 1977) that induces a perfect (Selten, 1975) and hence sequential equilibrium in every game with the same normal form. These properties assure that a prediction is immune to how the game is presented in normal form or as one of many equivalent extension forms.

c. [Forward Induction] A stable prediction contains a sequential equilibrium for which a player’s belief in each event assigns positive probability only to the other player’s strategies that are optimal replies in the continuation of the game, if some allowed the event to occur. This property assures that beliefs are also consistent with rationality. It implies popular criteria (such as the intuitive criterion, divinity, and ‘never weak best reply’) that exclude from beliefs those strategies that are not optimal in any weakly sequential equilibrium with the same outcome.

d. [Iterative Elimination] Some subset of a stable prediction is itself a stable prediction of the reduced game obtained after iterative elimination of weakly dominated strategies and strategies that are inferior replies at every equilibrium in the original stable prediction.

e. [Decomposition] The stable predictions of the product of two independent games are the products of their stable predictions.

f. [Player Splitting] A stable prediction is immune to splitting a player into agents who act in their assigned contingencies.

g. [Ordinality] Stability depends only on the ordinal properties of players’ preferences.

Example

I use an example to illustrate. Game theory is often used to study the role of signaling in markets. An influential study of signaling is by Michael Spence (1973). He supposes that a firm offers a higher initial salary to a worker with more credentials, such as school and university degrees, because the firm believes that more credentials indicate that the worker has greater ability and therefore will be more productive. Anticipating this, the worker obtains credentials to get higher salaries, and importantly, a worker with greater ability obtains more credentials because he or she can do so more easily. Thus the firm's beliefs are justified – a candidate who presents more credentials tends to have greater ability.

But signaling games have many equilibria. At one extreme, workers with different abilities obtain different credentials and different salaries; and at the other extreme, workers of all abilities obtain the same credentials and receive the same salaries. So, which equilibrium is predicted by the theorem? When a standard technical assumption (called "single crossing") is satisfied, the prediction that satisfies the axioms is the one in which workers with different abilities obtain different credentials and different salaries. This result is a consequence of the forward induction property.

Remarks

From the perspective of economic theory, the chief implication of the theorem is that an axiomatic characterization is possible (which some had doubted), although here we do so only for two-player games with perfect recall and generic payoffs.

From the perspective of social science more generally, the theorem is somewhat baffling. It has the pleasant implication that four simple axioms characterize those predictions that have all the desirable properties that have been sought over the past thirty-five years. But also it has the unpleasant implication that a prediction satisfying the axioms is necessarily a stable set – and identifying such a set requires solving a difficult computational problem. Some will be dismayed by the conclusion that the tools of algebraic topology are necessary to derive a prediction of rational strategic behavior.

For those who apply game theory to economic problems and those who use games for experimental and empirical studies, there are practical implications. One is a caution: if one predicts a game's outcome that does not result from equilibria in a stable set, then some axiom is violated. In particular, if the outcome results from a sequential equilibrium in undominated strategies, then this prediction depends on allowing players to be influenced by framing effects derived from some metagame that embeds the game being studied. Thus the prediction could be invalidated by changing the context within which the game is presented.

A secondary implication is that, after deviations from equilibrium play, the continuations of all equilibria in a stable prediction remain undominated and sequentially rational, where those that are not sequential equilibria of the original game are justified by beliefs induced by a quasi-perfect (hence sequential) equilibria of some corresponding metagame that embeds the given game. This resolves a conundrum posed by Reny (1992).

Conclusion

Given the wide application of game theory in economics, it is unfortunate that axiomatic foundations have been delayed so long. Ultimately one wants a theory of strategic behavior in multi-person situations that is based on elementary axioms, analogous to those for single-person decision theory. I see recent progress as advancing the program for axiomatic development posed by Kohlberg and Mertens in 1986. Much remains to be done – an arbitrary number of players, nongeneric payoffs, stochastic games, games with bounded recall, … .

I thank my hosts at the Toulouse School of Economics and the City of Toulouse for their generosity in inviting me to participate in this remembrance of Jean-Jacques Laffont. I hope my lecture advances their efforts to extend the beneficial influence of Jean-Jacques' contributions to the theory and practice of economics.