Cristina Bicchieri

Carol and Michael Lowenstein Professor
Director, Philosophy, Politics and Economics

University of Pennsylvania, USA

Why were you initially drawn to game theory?

I was a student in philosophy of science at Cambridge University in the early 80's. I was interested in Bayesian confirmation theory, but I was not happy about the formal tools available to answer questions about why we adopt a hypothesis or choose a theory. Those were the years in which Kuhn's ideas about scientific revolutions and sociological attempts to explain scientific practices were dominant in the community. I thought the social dimension of science was important, but I never believed we are dupes that respond automatically to the social environment surrounding us. Social influences should be incorporated in a model of choice, I thought, but how to proceed to do it was far less obvious. I wanted to show that there is a conventional element in the choice of which method or model to apply, but also model it as a rational choice. It was a choice, though, that did not occur in a vacuum: it had to depend upon what one expected other scientists to choose. Decision theory offered only a partial answer to my quest for a model of rational decision-making. It tells us how to make a rational choice against Nature, whereas I wanted to know what choosing rationally means when the outcome depends on what other people choose, too. Game theory gave me the answers I wanted. An article I published in 1988 summarized my views about how game theory should be applied to the study of scientific practices. It contained themes and ideas I have developed later on: social norms, conventions and common knowledge among them. At that point I had become less interested in how scientists make choices than in what it means to choose rationally in an interactive context, and whether rationality and common knowledge of rationality alone could guarantee that players coordinate upon an equilibrium. These are also important philosophical issues: epistemology contends with questions about knowledge, belief, and rationality, what they mean and how to model them. Game theory challenged philosophers to think in terms of interactive epistemology: what does it mean for a collective to have common beliefs or knowledge, and what the consequences of this knowledge are for the social outcomes resulting from agents' interactions. I eventually moved on to explore the relation between agents' knowledge and solutions to games. Though what drew me to game theory many years ago was a specific question about scientists' decision making, what draws me now is the recognition that philosophy cannot do without the language of game theory. In many ways, the two fields are interconnected, and can greatly benefit from each other.

What example(s) from your work (or the work of others) illustrates the use of game theory for foundational studies and/or applications?

There are many areas of philosophy that interact in a fruitful way with game theory. An important intersection between game theory and philosophy is the 'epistemic approach' to game theory. Epistemology traditionally studies concepts such as truth, justification, knowledge and belief. Game theory usually assumes agents have common knowledge (beliefs) of the structure of the game and their mutual rationality. However, it took time before game theorists recognized that it is important to explicitly formalize the hypotheses we make about the knowledge and beliefs of the players. At the beginning of the subject, the fact that decision theory had clear foundations (for example, Savage's axioms) seemed sufficient. However, decision theory treats the agents' probabilistic beliefs as exogenous, whereas in game theory the main source of uncertainty for an agent is the way other agents will behave. If we can infer that a rational agent will behave in a particular way, then another rational agent should also predict the first agent's behavior. Probabilistic beliefs necessarily become endogenous. The epistemic approach to game theory provides a formal analysis of strategic reasoning, making explicit players' knowledge (or beliefs) about the structure of the game and the strategies, knowledge and beliefs of other players. It also has the important merit of providing an epistemic foundation for solution concepts. For example, given a certain family of games G, there will be some strategy profiles S compatible with the knowledge and beliefs attributed to the players. We can thus show that a solution S(G) captures the epistemic hypotheses E that, in turn, yield that solution.

A case in point is the backwards induction paradox. The ‘paradox’ arises thus: In the standard model which admits of backwards induction arguments, to determine the solution we have to begin by saying how rational players would behave at all penultimate nodes. We assume common knowledge of players’ rationality, and infer that they will never get to all but one of these. Hence, since we assume that rational players know our theorem, for any one n of these nodes, if a player arrived at n she would know that someone is not rational. But common knowledge of players’ rationality implies simple knowledge of players’ rationality. So at this node the player would know inconsistent propositions. Many authors such as Rosenthal, Binmore and Reny had tried to explain what is paradoxical in these problems; my contribution was to explicitly introduce into the game the knowledge that players have about each other (hence my dubbing them knowledge-dependent games), and show that limited knowledge, but not common knowledge, supports the backwards induction solution. For example, if players have only mutual knowledge of rationality of a certain limited degree, then a player deliberating at the start cannot infer from her knowledge (because the inference would requires knowledge of more than this limited degree) that the player at the penultimate node would, finding himself there, be bewildered.

More generally, I argued (1989, 1992) that the backwards induction solution of a game is a knowledge-consistent play of a knowledge-dependent game. The reverse is also true. Thus there is an isomorphism between backwards induction and knowledge-consistent plays of certain associated games. The important point I made is that some knowledge-dependent games have no knowledge-consistent plays. That is, if we incorrectly translate a game into the associated knowledge-dependent game, some player will be unable to reason about the others well enough (or consistently enough) to infer which action she should choose. Too little knowledge impairs deduction of the proper action, but too much knowledge is equally damaging.

Another important application of the epistemic approach is the treatment of belief revision in games. In games of imperfect information, an agent reaching an information set which had zero probability under the equilibrium strategies must have some belief about what happened in the past. Since Bayes' Rule is not applicable to updating after zero probability events, there is no obvious 'rational way' for doing this. Yet beliefs at such information sets are crucial for determining equilibrium play. To play an equilibrium, a player must know what would happen at off-equilibrium information sets. This is a crucial problem in the refinements literature, where an equilibrium is rejected as implausible if it is unstable with respect to small deviations. However, all depends upon how a deviation is interpreted by the players. A deviation is, from the viewpoint of playing a given equilibrium, a contrary to fact event. Philosophers' work on counterfactuals and belief revision is clearly important here. In particular, the work of Gardenfors and Levi on minimum loss of information criteria allows for a ranking of refinements that is much more plausible than the host of intuitive arguments provided by game theorists to justify some equilibria as more 'reasonable' than others. Since minimum loss of information is not just a quantitative, but also a qualitative criterion, I argued (1988) that interpreting deviations as mistakes, as opposed to rational signals (whenever this interpretation makes sense), deprives rationality of its explanatory and predictive power, and therefore causes a greater loss of information.

The search for better models of agents' reasoning includes finding ways to formalize the theory of the game (T) that is used by an agent to infer its moves and therefore to compute a solution to the game. Typically, T might be formalized in a classical, first-order logic. Such a logic is sufficient to represent the structure of the game and the associated payoffs, and to infer an optimal sequence of moves relative to a given utility function. If one wants to represent the players' reasoning processes and beliefs, then it becomes necessary to use modal logics. These are, however, monotonic logics, in that any proposition p entailed by a set of axioms A remains entailed by any set of axioms B that includes A.

Using a monotonic logic leads to problems whenever we want to model the possibility of unexpected events occurring. For example, if an agent's theory predicts that another agent will make a certain sequence of moves, the other agent choosing otherwise is an unexpected occurrence. Such an event contradicts some of T's premises. If T is expressed in a monotonic logic, this theory, when augmented with statements to the fact that unexpected moves have occurred, becomes inconsistent. A more realistic theory of the game should be a theory that is robust in the face of deviations from predicted outcomes, i.e. it must be a theory that allows agents to play even in the presence of deviations. Antonelli and I (1995), used Reiter's default logic to formalize agents' reasoning. We specified a default first-order theory (W,D) for generic, finite extensive form games with perfect recall. Such a theory comprises two main modules or parts. The first part of the theory describes a mechanical procedure π^* that computes the set of undominated paths through a finite tree representing a game of perfect or imperfect information in extensive form. The second part contains a set of first-order axioms W, and a set of defaults D. W includes a description of the structure of the game and behavioral axioms specifying that whenever a non-terminal node (or information set) is reached, an agent will choose exactly one among the possible moves. The defaults D represent defeasible behavioral principles to the effect that agents only choose moves allowed by recursive application of π^*. We assumed W to contain Primitive Recursive Arithmetic, which is necessary to define a function π^*representing a particular computing (pruning) procedure. The function π^* takes a set of nodes (or information sets) as input, and returns a set of paths through these nodes as output. In our work, we introduced a particular π^* function, embodying a specific procedure for recursively pruning the tree. In general, one may want to employ different procedures on different occasions or for different purposes. Only the first part of the theory (specifying π^*) would have to be changed, leaving the behavioral axioms and the defaults unchanged.

The procedure π^* that allows agents to recursively compute their own and other agents' undominated paths and information sets throughout the entire tree embodies a particular rationality principle: a rational player is one that only plays admissible strategies, where an admissible strategy is one that is not weakly dominated. Recursive application of π^* along the tree embodies the concept of iterated elimination of weakly dominated strategies. Our behavioral axioms did not contain an explicit definition of rationality. However, such an assumption is already implicitly made in attributing to players the capability of computing $\pi^*$and choosing according to it along the tree. When recursive application of π^* returns a unique path as the solution of the game, we proved that the solution corresponds to a Nash equilibrium. Moreover, the solution concept we proposed rules out all those Nash equilibria that contain weakly dominated strategies. Building on the earlier work with Antonelli, Oliver Schulte and I (1997) proposed a new, more general version of the pruning procedure, one that provides a formal definition of agents' common reasoning about admissibility. We obtained several interesting results:

- Our definition of common reasoning about admissibility coincides with order-free elimination of weakly dominated strategies in the strategic form;
- In the extensive form, a strategy may prescribe choices in parts of the tree that will never be reached if that strategy is played. If we evaluate strategies only with respect to information sets that are consistent with them (i.e., those that can be reached if the strategy is played), we are led to the concept of sequential proper admissibility. A strategy is sequentially properly admissible in a game tree just in case it is admissible at each information set consistent with the strategy. The strategies that are consistent with common reasoning about sequential proper admissibility in the extensive form are exactly those that are consistent with common reasoning about admissibility in the strategic form representation of the game. Thus the solution given by common reasoning about admissibility does not depend on how the strategic situation is represented.
- We defined a credible forward induction signal as a signal consistent with common reasoning about sequential admissibility. If we allow agents to consider only credible signals, common reasoning about sequential admissibility yields typical forward induction solutions in games of imperfect information.
- In games of perfect information, common reasoning about sequential admissibility yields typical backward induction solutions. Note that the recursive pruning procedure does not start at the final nodes. Our procedure allows agents to consider the game tree as a whole and start eliminating branches anywhere in the tree by applying iterated admissibility, and therefore it does not follow Zermelo's backward induction algorithm. For example, suppose that in a game tree a move m at the root is strictly dominated by another move m^0 at the root for the first player. Our procedure rules out m immediately, but the backward induction algorithm eliminates moves at the root only at the last iteration.

One advantage of this approach is that of providing a unified treatment of several solution concepts that were previously held to be different, if not incompatible. Thus a unique mechanical procedure that embodies common reasoning about admissibility can be applied in a wide variety of games. Another advantage is that rationality, and common reasoning about rationality, need not be explicitly defined. They are embedded in the mechanical procedure an agent is provided with. Much work remains to be done in this area, especially important for applications to distributed AI, where keeping the procedure and the axioms separated may present an advantage.

What is the proper role of game theory in relation to other disciplines?

Game theory, though it is extensively used in economics and other social sciences, as well as in computer science, biology and philosophy, is an autonomous discipline. It applies to all situations in which decision makers interact and the outcome depends on what the parties jointly do. Decision makers may be people, firms, political parties, animals, robots and even genes. When firms compete for market share, politicians compete for votes, jury members have to decide on a verdict, animals fight over prey or genes compete for survival, we have a strategic interaction. Game theory is the formal language in which we model what all these interactions have in common. Yet it would be wrong to think that, since it is similar to a formal language, game theory only lends itself to a precise, skeletal representation of properties that are already there, though expressed in a less formal way. The role of game theory is that of a model: it gives us an idealized version of the phenomena we study, but it also leads us to explore particular facets of such phenomena.

Using a model, even when it is a formal model (as opposed to, say, a physical one), brings about new inferences, suggests new properties and in a sense changes the thing or process that we model. The kind of tools game theory gives us are apt to change the way we understand the phenomena we model with those tools.

As an example, think of the role game-theoretic models are playing in ethics and political philosophy. These disciplines deal with moral rules, social contracts, conventions of justice and the like, all concepts that can be given a precise meaning as equilibria of repeated games. Non-cooperative game theory is an invaluable tool that has been little understood in philosophy. And so is evolutionary game theory.

Brian Skyrms, for example, has done seminal work showing how moral norms can evolve, and what their subsequent dynamics might be. Ken Binmore, Robert Sugden, Peter Vanderschraaf, Jason Alexander and myself have applied game theory to the evolution of norms, as well as to more classical problems in political philosophy. David Lewis and Edna Ullman-Margalit were the first to see the potential of game theory to explain conventions and to differentiate them from other rules, as well as to link game theory with the philosophy of language. In fact, a new and exciting area of application of game theory to philosophy is the study of how meaning can emerge. In all these areas, game theory has helped to sharpen our intuitions, allowing for a 'rational reconstructions' of difficult concepts and an explanation of how social contracts, norms, conventions, values and even meaning can emerge out of various forms of interactions among agents who did not plan or expect such results.

Finally, let me mention the results that experimental game theory brings to bear on the development of ethical theories. Experiments on Ultimatum, Dictator, Trust and Social Dilemma games are helping us understand how people form fairness judgments, the cognitive dynamics involved in the process, and what drives 'fair' behavior on one occasion and dampens it in another. These are important steps that any philosopher should take in the direction of building better normative theories. Naturalizing ethics does not mean reducing what ought to be done to what is in fact done: this would be a trivial naturalistic fallacy misstep. What instead needs to be done is build our normative theories upon the solid foundation of what we know individuals can in fact do, and this is a whole different project. I have embarked on this project long ago, by trying to show that our ethical norms are just collectively defined and supported social norms. Some such norms are more entrenched that others, but the cognitive processes underlying norm-following, and the biases we all face in filtering and processing the social information that will ultimately decide whether or not we act in a pro-social way, are essentially the same. Without knowledge of such cognitive processes, and the behaviors they engender, ethics is condemned to remain an abstract and fairly useless endeavor.

As an example of the multifaceted use of game theory in developing better philosophical theories, consider building a theory of social norms (Bicchieri 2006). In order to provide a testable, operational definition of social norms, one has to define them in term of conditional preferences and beliefs, and show that norms, when they exist and are followed, transform a mixed-motive game such as a prisoner's dilemma or a trust game into a coordination game of which the norm is a salient equilibrium. When we encounter a new situation, we must decide whether to obey the norm or act in a selfish way. It is as if we are playing a Bayesian game in which we assess the probabilities that the opponent is selfish or norm abiding. The theory of norms I propose predicts that, if the right kinds of expectation are present, most subjects will follow whatever norm is relevant to the decision situation. We can set up behavioral experiments in which we manipulate expectations and test this hypothesis. Clearly a game-theoretic model shapes the way we address these questions, and directs us to interesting new solutions.

What are the most important open problems in game theory and what are the prospects for progress?

There are several areas of game theory, such as cooperative games, that have languished for some time. Here however I want to concentrate on the applications I just mentioned, since a lot more need to be done in these areas.

DAI applications. The epistemic approach is crucial in applications to distributed artificial intelligence (DAI). DAI focuses on solutions of problems by a multi-agent community, such as distributed planning systems, or Web agents that retrieve information for their users. Since agent performance is more effective if interactions with other agents involve coordination or cooperation, any designer of agents that act in a multi-agent environment faces the problem of encoding a strategy of interaction with other agents. Traditionally game theory has not been concerned with agents' design, and only relatively recently has explicitly dealt with formal models of agents' strategic reasoning. Thus, even if game theory can be extremely useful in providing us with methods for proving properties that are useful to adopt for designing agents, there is still a lot of work left in order to adapt these methods to the design of artificial agents. There is a need to focus on the reasoning processes of the individual players rather than on the framework within which their encounters take place. When adapting game theoretic models to a DAI environment, the choice of a strategic model applicable to the specific DAI problem must include, among other things, the development of techniques for searching for appropriate strategies that will enable agents to reach an equilibrium. Artificial agents should be programmed; mere identification of a solution is not sufficient. To derive a solution an artificial agent must possess reasoning capabilities, an algorithm that will determine its strategic behavior given the information the agent is endowed with. If agents use their available information to reach a conclusion as to how to play, this information must be explicitly represented. Often even the simplest strategic interaction has several possible solutions (equilibria). If agents are the product of different software designers who do not share a common protocol, giving each agent the ability to reason to a solution, and heuristic rules to interpret other agents' behavior, becomes imperative. Moreover, when dealing with artificial agents, the complexity of deriving a solution becomes both measurable and crucial; the computationally intractable (or at least impractical) assumptions of omniscience and common knowledge (Parikh 1987, 1995) must be relaxed and replaced with more realistic, implementable assumptions.

Experiments. Experimental game theory has seen a remarkable growth in recent years. Experiments show that the usual auxiliary hypotheses about agents' selfish motives have to be changed, at least in many cases in which pro-social behavior is involved. Unfortunately, though many new utility functions have been proposed to explain what we observe in experiments, we still lack utility functions general enough to subsume a variety of results and specific enough to allow for meaningful predictions. Moreover, there are many open questions about the role of emotions in decision-making, and their relation to beliefs and expectations. New research done in neuroeconomics might shed light on these issues, but I believe we still need behavioral studies to assess the role of expectations, measure them, and see how manipulating expectations may lead to dramatic behavioral changes.

Evolutionary models. In this area, too, there is a lot of work to be done. Traditional replicator dynamics models are not adequate to model cultural evolution, and we need to develop more sophisticated imitation/learning models that take into account psychological factors. I expect more work will be done in integrating the results of lab experiments into better, more realistic evolutionary models. For example, endowing agents with utility functions that represent more accurately their motivations will allow us to build evolutionary models of, say, the emergence of institutions such as social norms that have greater explanatory value.