



BA 513/STA 234: Ph.D. Seminar on Choice Theory
Professor Robert Nau
Fall Semester 2004

Readings for class #9: Criticism of game theory

Primary readings:

- 1a. "Game theory and its discontents," chapter 8 of *Prisoner's Dilemma* by William Poundstone, 1992
- 1b. "Subjective probability and the theory of games" by Jay Kadane and Patrick Larkey, *Management Science*, 1982 (with comment by John Harsanyi, reply by Kadane and Larkey, and rejoinder by Harsanyi)
- 1c. "Rational choice: contributions from economics and philosophy" by Robert Sugden, *Economic Journal*, 1991
- 1d. "Explaining Everything, Explaining Nothing? Game Theoretic Models in Industrial Economics" by John Sutton, *European Economic Review*, May 1990

Supplementary readings:

- 2a. "The confusion of IS and OUGHT" by Jay Kadane and Patrick Larkey, *Management Science*, 1983 (with comment by Martin Shubik)
- 2b. "Comments on the interpretation of game theory" by Ariel Rubinstein, *Econometrica*, 1991
- 2c. "The problems of game theory," chapters 5 and 6 from *Game Theory and Economic Modelling* by David Kreps

Guide to the readings:

Poundstone's book is a history of the early years of game theory intertwined with a biography of John von Neumann. This chapter shows that by the early 1950's many of the shortcomings of game theory that are still debated today were already well understood, and a backlash had already started to occur. The quotation from Gregory Bateson's 1952 letter to Norbert Wiener calls attention to the problem of starting from the assumption that the rules of the game are already fixed and commonly known: "What applications of the theory of games do, is to reinforce the players' acceptance of the rules and competitive premises, and therefore make it

more and more difficult to conceive that there might be other ways of meeting and dealing with each other...”

The lament by RAND’s John Williams in 1954 describes a point of view that is still echoed by contemporary rational choice critics: “[game theorists] are often viewed by the professional students of man as precocious children who, not appreciating the true complexity of man and his works, wander in wide-eyed innocence, expecting that their toy weapons will slay live dragons just as well as they did inanimate ones.” The strategy of “mutually assured destruction” that was formulated in those years, whose legacy of nuclear weapon stockpiles is a live dragon that troubles us today, was in part the brainchild of game theorists studying simple 2×2 games like “chicken.”

As Poundstone observes, game theory has a Machiavellian flavor. The rules of the game are supposed to be expressed in units of personal utility. Once the utility values have been assigned to outcomes of the game, no moral reasoning is required to decide how to play: only mathematical calculations remain to be performed. As such, game theory itself is objective and value neutral—only the utilities are subjective. But among other problems, those utilities are hard to measure with the precision that the theory requires, particularly in games that do not have equilibria in pure strategies (a.k.a. “saddle points”).

Poundstone next recounts some of the early experiments on repeated prisoner’s dilemma (and non-dilemma!) games at Ohio State, in which some players repeatedly defected even when it was not in their own interest, seemingly out of a desire to merely beat their opponents, while other, more civic-minded players cooperated even when it was not in their own best interest. Poundstone concludes: “About the only believable conclusion of these studies is that those inclined to cooperate in one context usually do so in other contexts. Some people are habitual cooperators, and some are habitual defectors.” More recent studies have shown that, indeed, students of *economics* are some of the worst habitual defectors (Robert Frank et al., “Does studying economics inhibit cooperation?” *Journal of Economic Perspectives*, 1993)

These results highlight an intrinsic dilemma of game theory. Ideally we would like to be able to discuss a game in purely abstract terms, as if any rational person should be expected to play it the same way. But realistically, even in the simplest parlor games or laboratory experiments, the psychological baggage that the subject brings to the gaming table may strongly affect her style of play. We can, of course, try to rationalize such results by allowing that, when presented with the same “objective” game, different individuals may assign different subjective utilities to the outcomes—in some cases behaving as if their own utility depends on fates suffered by others as well as themselves—but this rationalization is not very helpful unless we (and other players) can predict how the individuals will make those utility assignments.

Kadane and Larkey’s 1982 paper has been widely cited by both detractors and defenders of game theory. (For example, it is mentioned by Aumann in his 1987 survey paper.) Legend has it that this paper was initially rejected by *Management Science* because it received negative reviews from the two initial referees. The authors went back to the editor and pointed out that one referee recommended that the paper be rejected because its claims were obvious, while the other recommended that it be rejected because its claims were preposterous. This divergence of

opinion suggested that they had struck a sensitive nerve in the field. The paper was eventually published, and Harsanyi was invited to write a response. A number of other authors later weighed in with their own comments, published in subsequent issues of the journal in the same year.

Kadane and Larkey open their paper with a comment on the curious gulf that exists between the two long-sundered branches of decision theory that emanated from von Neumann and Morgenstern's pioneering work: *game theory* (as subsequently developed by Nash, Harsanyi, and Selten) and *subjective probability theory* (as subsequently developed by Savage, incorporating ideas of Ramsey and de Finetti). Kadane and Larkey point out that from a subjectivist, Bayesian perspective, the decision maker needs only take into account his own first-order beliefs about his opponent's actions at the time he chooses his own strategy, rather than getting caught up in an infinite regress. As they point out: "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 own 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." They do admit that repeated games leave more room for higher-order beliefs to matter, insofar as they affect how the individual might learn over time (notwithstanding Hacking's criticism of the very concept of "Bayesian learning"!).

Harsanyi thunders back that "Kadane and Larkey ... do not seem to realize that their approach would amount to *throwing away essential information*, viz., the assumption (even in cases where this is a realistic assumption) that the players will act rationally and will also *expect* each other to act rationally. Indeed, their approach would trivialize game theory by depriving it of its most interesting problem, that of how to translate the intuitive assumption of mutually expected rationality into mathematically precise behavior terms (solution concepts)."

What is interesting about this exchange is that Harsanyi, unlike many game theorists, actually characterizes himself as a Bayesian. The gulf that separates Kadane and Larkey from Harsanyi is the gulf between "subjective" Bayesians in the tradition of Savage (who are generally comfortable allowing people to believe whatever they want, as long as their beliefs are consistent) and "necessitarian" Bayesians in the tradition of Harold Jeffreys and E.T. Jaynes (who feel that prior beliefs should be objectively determined by prior information). As Kadane and Larkey point out (and Harsanyi basically agrees), "*solution concepts [of game theory] are a basis for particular prior distributions.*"

Harsanyi certainly has a point: it is possible that the structure of the payoffs in a game might lead reasonable people to form similar beliefs about how their opponent is likely to play, just as the physical symmetry of a random device such as a six-sided die or a roulette wheel might lead reasonable people to agree on the probabilities in a game of chance, even in the absence of empirical frequency data. Some games *do* have obvious solutions. But the problem is that, as we have seen, many games either do *not* have obvious solutions or else the obvious solutions violate the solution concepts that theorists would like to impose. The search for game theory's holy grail—a universal solution concept that would uniquely and uncontroversially determine the

outcome of any game—just did not pan out, despite the heroic efforts of Harsanyi, Selten, and others. This leaves us (almost) back at square one: which solution concept do you expect your opponent to use, and which of several equilibria allowed by the solution concept will he then play (and expect you to play in return, and so on), not to mention the problem of knowing what rules your opponent is really playing by!

In their reply to Harsanyi, Kadane and Larkey hold out a sort of olive branch: they agree with Harsanyi's statement that "what we need is an empirically supported *psychological* theory making at least probabilistic predictions about the strategies people are likely to use ... given the nature of the game and given their own psychological makeup." But Harsanyi rejects their overture in his rejoinder, claiming that "the question of how to act against highly *rational* opponents ... actually is, and has always been, the main intellectual attraction of game theory." About the latter claim there can be little argument: the attempt to solve the riddle of reciprocal expectations of rationality is what puts the fun in game theory.

In their followup 1983 paper, Kadane and Larkey, attempt to seize even higher moral ground, calling attention to the confusion between descriptive and normative modeling—i.e., between what "is" and what "ought" to be—that is prevalent in game theory and much of the rest of microeconomics. Off come the gloves. They state:

The confusion between "is" and "ought" in game theory is widespread and is a serious obstacle to developing the theory along more productive lines. A distressingly large proportion of the research in the social, behavioral, and management sciences can be categorized as either cumulatively useless or noncumulatively useless. There are modes of theorizing that have not and are never likely to lead to useful prescriptions or useful predictions of human behavior. Game theory epitomizes the cumulatively useless. It is astounding that thirty-plus years of (often elegant mathematical) theorizing and experimentation could produce so little of value in instructing people on how they should behave in conflict situations and in predicting how they *do* behave in conflict situations.

To illustrate this viewpoint, they present an example of a two-player zero-sum game in which there is already some history of repeated play, and point out the difficulty of either prescribing what one of the players should do next or in predicting what both are likely to do next. (By the way, recent work in experimental game theory has focused on similar situations, using models of adaptive learning to explain patterns in the data.)

Kadane and Larkey conclude that a subjective, Bayesian perspective provides the correct solution. In order to *prescribe* rational behavior for yourself, you first need a *predictive* model of your opponent's behavior. The latter predictive model may, in some cases, be based on the attribution of "rules of thumb" to your opponent, including ideas such as playing minimax strategies in zero-sum games. But, they add: "At best, this rule-of-thumb is a partial basis for forming your prior about your opponent's likely behavior in certain game situations. It is not a logically compelling prescription for your own play. And it is not a very accurate predictive theory for most games." The same asymmetrically descriptive/prescriptive view of interactive decision making is advocated by Howard Raiffa (*The Art and Science of Negotiation*, 1982) and

Max Bazerman & Maggie Neale (*Negotiating Rationally*, 1992) as a rational basis for bargaining and negotiation.

Here too, Kadane and Larkey's paper was accompanied by a response from an eminent game theorist, in this case Martin Shubik. Shubik points out some successes of cooperative and noncooperative game theory under conditions where their respective assumptions about information and communication happen to apply, but he is generally more sympathetic to Kadane and Larkey's position than was Harsanyi. (In a paper entitled "What is an application and when is a theory a waste of time" that was published several years later in *Management Science*, Shubik discussed the distinctions among "high church game theory," "low church game theory," and "conversational game theory," acknowledging that the latter two forms of game theory are often more useful than the former.)

Sugden's very thoughtful paper surveys a number of foundational issues in rational choice theory, some of which involve Savage's theory of subjective expected utility and others of which involve game theory. Concerning SEU theory, Sugden echoes some of the questions raised earlier by Shafer about the normative status of axioms such as completeness and transitivity and about the concept of a "consequence" that plays such an important role in the theory. In this context, he mentions some of the work on *regret theory*, of which he was one of the principal authors. Concerning game theory, he notes that Savage's axioms are seemingly inapplicable to game situations in which the objects of the players' uncertainty are each others' strategies.

"For Savage, the description of an event can make no reference to any act, but the 'event' that one player's opponent plays a particular strategy in a particular game cannot be described without reference to the game itself, and hence to the first player's set of feasible acts. Nevertheless, it is standard practice in game theory to use expected utility theory, and even to appeal to Savage for intellectual support in doing so. This practice leads to serious problems..."

Game theory's assumption of *common knowledge of rationality* requires a player to imagine that her reasoning is transparent to her opponents, and vice versa, in which case her beliefs about events (namely, her opponent's strategies) are not independent of her own strategy choice. Sugden goes on to criticize Harsanyi's and Aumann's use of the common prior assumption, observing that "it is hard to see how this position can be compatible with Savage's subjective conception of probability." He also calls attention to other problems in the assumption of common knowledge of rationality: in some games this assumption is insufficient to lead to equilibrium, and in others it appears to be incoherent. For example, in the famous centipede game, backward induction leads to the self-defeating strategy of terminating the game on the first move—but it has no answer to the question of what will happen if a player chooses to continue the game instead. Hence, strictly speaking, we cannot say that a player achieves a higher expected utility by ending the game than by continuing: the expected utility of continuing is undetermined. (Various refinements of Nash equilibrium also get mired in the question of having to explain what happens when a player deviates from the equilibrium path, even though the players are assumed believe that a deviation has probability zero.) Other puzzles arise in simple games of coordination and games of commitment: rationality can undermine the players' efforts to coordinate on a "salient" solution of a game, and in some situations a player can even

be better off by behaving irrationally. Sugden concludes: “There was a time, not long ago, when the foundations of rational choice theory appeared firm, and when the job of the economic theorist seemed to be one of drawing out the often complex implications of a fairly simple and uncontroversial set of axioms. But it is increasingly becoming clear that these foundations are less secure than we thought, and that they need to be examined and perhaps rebuilt.”

Rubinstein’s paper is not an attack on game theory, but an attempt to reinterpret it in a more realistic fashion. (Rubinstein is an eminent game theorist, the architect of a celebrated model of noncooperative bargaining.) He argues that a game model should be interpreted as a representation of the players’ *perception* of their situation, rather than “physical rules of the world,” and he suggests that the concept of a player’s “strategy” in a game should be broadened to include “those considerations which support the optimality of his plan rather than merely ... a ‘plan of action.’” This interpretation blurs the distinction between a “solution” and “solution concept” to some extent. On this view, “a strategy encompasses not only the players’ plan but also his opponents’ beliefs in the event he does not follow that plan.” Rubinstein next takes up the question of the interpretation of mixed strategy equilibria. He mentions Harsanyi’s concept of “purification,” in which mixed strategies are viewed as pure strategies of games with exogenous, payoff-irrelevant noise, but he notes that this interpretation, though consistent, is fraught with problems. He seems to prefer Aumann’s interpretation of a mixed strategy as a description of the joint beliefs held by all *other* players with respect to the strategy of a given player, but he notes that “this renders meaningless any comparative statics or welfare analysis of the mixed strategy equilibrium and brings into question the enormous economic literature which uses mixed strategy equilibrium.”

Rubinstein also discusses the famous “burning money” equilibrium concocted by Eric von Damme. This example astonished the game theory community when it was first presented in the late 1980’s—indeed, at the time, many felt it was a catastrophe for the refinements-of-Nash-equilibrium movement. Consider the familiar battle-of-sexes game. As we have seen, this game has two efficient—but unequal—pure strategy Nash equilibria: it is a simple game of coordination in which neither player has an intrinsic advantage and domination arguments do not apply. The problem is how to decide which player gets his or her best outcome and which has to settle for second-best. Von Damme demonstrated that if one player is imagined to have the option of burning a dollar at the outset of the game, then the iterative deletion of weakly dominated strategies in the augmented game leads to a unique equilibrium in which that player is awarded her best pure-strategy outcome without actually burning money. Somehow the threat—even the imaginary threat—of *self-abuse* provides a strategic weapon that can be used to force one’s opponent to settle for second-best. Marvelous! Rubinstein argues that it is necessary to impose “relevancy” conditions in order to eliminate absurdities like this. (An alternative view is that the *iterative* deletion of weakly dominated strategies is simply an incoherent solution concept. Correlated equilibrium yields a very sensible solution to this game.)

Rubinstein’s concluding paragraph seems to echo some of the points raised by Kadane and Larkey and by critics of the rational choice paradigm such as Herbert Simon:

“There exists a widespread myth in game theory, that it is possible to achieve a miraculous prediction regarding the outcome of interaction among human beings using

only data on the order of events, combined with a description of the players' preferences over the feasible outcomes of the situation. For forty years, game theory has searched for the grand solution which would accomplish this task. The mystical and vague word 'rationality' is used to fuel our hopes of achieving this goal. I fail to see any possibility of this being accomplished. Overall, game theory accomplishes two tasks: It builds models based on intuition and uses deductive arguments based on mathematical knowledge. Deductive arguments cannot by themselves be used to discover truths about the world. Missing are data describing the process of reasoning adopted by the players when they analyze a game. Thus, if a game in the formal sense has any coherent interpretation, it has to be understood to include explicit data on the players' reasoning processes. Alternatively, we should add more detail to the description of these reasoning procedures. We are attracted to game theory because it deals with the mind. Incorporating psychological elements which distinguish our minds from machines will make game theory even more exciting and certainly more meaningful."

We will return to the question of inductive versus deductive reasoning in a few weeks. But in the meantime, rational choice modelers take heed: game theory does not yield miraculous predictions based (only) on the players' preferences!

The last two chapters of **Kreps'** book range over a similar set of problems in game theory and come up with similar recommendations: game theory needs to consider the actual decision processes of the players and to incorporate concepts of bounded rationality in order to be more realistic. Kreps points out the problems posed by the existence of multiple equilibria, by the very concept of "equilibrium," by the counterfactual reasoning used in refinements, by knowledge of the rules of the game, etc. His concluding chapter points to ways in which models of learning ("behavioral dynamics") might be used to overcome some of these problems. He also endorses the "new institutionalist" view that the rules of games that arise in society are themselves equilibria of more fundamental games played among boundedly rational individuals who are trying to economize on transaction costs. He ends on the following cautionary-but-hopeful note, which has by-now familiar echoes:

"Non-cooperative game theory ... has had a great run in economics over the past decade or two. It has brought a fairly flexible language to many issues, together with a collection of notions of 'similarity' that has allowed economists to move insights from one context to another and to probe the reach of those insights. But too often it, and in particular equilibrium analysis, gets taken too seriously at levels where its current behavioral assumptions are inappropriate. We (economic theorists and economists more broadly) need to keep a better sense of proportion about when and how to use it. And we (economic and game theorists) would do well to see what can be done about developing formally that sense of proportion.

It doesn't seem to me that we can develop that sense of proportion without going back to some of the behavioral assumptions we make and reconsidering how we model the actions of individuals in a complex and dynamic world. This means confronting some of the most stubbornly intractable problems of economic theory. But the confrontation has begun—indeed it is gathering steam—and so while I think we can be satisfied with some

of what has been achieved with these tools, it is appropriate to be happily dissatisfied overall; dissatisfied with our very primitive knowledge about some very important things and happy that progress is being made.”

Needless to say, his warnings to economists about the potential abuses of equilibrium analysis and the necessity of keeping a sense of proportion apply even more forcefully to applications of game theory outside of economics.