Editor: **Michael Wooldridge,** University of Liverpool, mjw@liverpool.ac.uk

The Triumph of Rationality

Michael Wooldridge, ***University of Liverpool***

Over the past decade, concepts and techniques from game theory have been both influential and successful in AI—and indeed, in computer science generally.

Artificial intelligence is, in a sense, all about decision making. After all, if you tend to make good decisions (however you measure "good"), then we might well be inclined to call you intelligent. But our decisions are rarely independent of those of others: to successfully operate as human beings in the everyday world, we must make decisions knowing that other individuals are also making decisions that will affect us and will have some bearing on the outcome of our decisions. Intelligent behavior requires us to take these other individuals into account, in particular the fact that they will make decisions based on their reasoning about how we will make decisions. If we want to build computer programs that can operate in such multiagent domains, it is natural to look for models and theories of such multiagent decision making that we can perhaps adapt for our purposes.

Since the mid 1990s, *game theory* has increasingly been studied in AI, and computer science generally, as a collection of models and concepts that can be used to investigate such multiagent decision making. Game theory originated in the first half of the 20th century from the study of games such as poker. (See the sidebar "The Founders of Game Theory.") Contemporary game theory, however, has much wider scope than simply understanding recreational games. We can understand game theory as an attempt to provide a mathematical theory through which to understand interactions in settings where the decision-makers are self-interested. This article provides a short introduction to some of the key ideas and concepts from game theory. Future articles will explore some of the applications of these ideas in AI and computer science.

What Is a Game?

Before we go any further, let us skewer a common misconception about game theory. Game theory is not just concerned with recreational games such as chess and poker—it is the theory of rational interaction between self-interested agents. Settings in which self-interested agents interact are ubiquitous in the real world, from global superpowers negotiating nuclear arms reduction treaties to you buying a second-hand iPod on eBay to you negotiating with your spouse about who should do the washing up. In this sense, the term "game" is a little unfortunate, because it seems to imply something trivial, of little consequence.

When game theorists use the term "game," they mean a mathematical model of a scenario in which self-interested agents interact and that captures all the information that is available to make a decision and leaves out all detail that is not germane to the decisions. There are many models of games in the literature, but the simplest model (noncooperative games in strategic form) includes the following components:

- a set of agents, which we call the players of the game;
- a set of choices for each player, called *strategies* for historical reasons, representing the possible ways that the player can choose to act—think of these as the "moves" of the game;
- a set of possible outcomes for the game—all the ways that the game could pan out;
- a description of which outcome will result for every combination of choices made by players in the game; and
- a description of each player's preferences over outcomes, which captures the player's self-interest. My preferences might well be different from yours, and yet the outcome depends on both our choices.

In a sense, the core problem of game theory is understanding what the right thing for a player to

1541-1672/12/$31.00 © 2012 IEEE
Published by the IEEE Computer Society

The Founders of Game Theory

It is a curious quirk of history that several of the most influential researchers in the early days of computer science and AI were also influential in the foundations of game theory. The foundations of contemporary game theory are generally traced to the 1944 publication of John von Neumann and Oskar Morgenstern's *Theory of Games and Economic Behavior* (see the "Further Reading" sidebar for details on this and other works described here). The Hungarian polymath John von Neumann (1903–1957) was, of course, extremely influential in the foundations of computer science. He was a key contributor to the architecture of modern-day computers, and we still refer to this architecture as the "von Neumann architecture." In 1928, von Neumann proved the first key result in game theory—the Minimax Theorem—and this remains one of the fundamentals of game theory to this day. The Minimax Theorem is, in a sense, the centerpiece of von Neumann and Morgenstern's lengthy masterwork, but it also demonstrates the limits of this early work, as it is essentially applicable only to a limited class of games, two-person zero-sum games.

A decade after the publication of the *Theory of Games and Economic Behavior* came the work of troubled genius John Forbes Nash, Jr. He formulated the solution concept that we now call Nash equilibrium. He also proved the key result, which states that every finite game has a Nash equilibrium if we permit randomization over strategies ("mixed" strategies). Sadly, after proving his key results, Nash suffered from serious mental health problems, which dogged him for several decades. Fortunately, the story has a happy ending. In 1994, Nash was awarded a Nobel Prize for his work, and his life was immortalized in the 2001 Hollywood movie *A Beautiful Mind*. Interestingly, Nash also had connections to the early days of AI. He was one of the 50-odd people invited to the 1956 Dartmouth Conference organized by John McCarthy, the event to which we traditionally date the founding of the modern discipline of AI.

Alan Turing, the founder of contemporary theoretical computer science and the author of arguably the first serious research paper on AI, was also apparently interested in computers and games. He wrote a chess-playing program in 1948, but, lacking a computer to run the program on, was limited to simulating its execution on paper.

do is. At one level, the answer is obvious: I should do whatever leads to the outcome I would be happiest with, according to my preferences. But there are two difficulties with this:

- In general, the outcome will not depend just on my choices, but on the choices of all the players in the game. When deciding how I should act, I should therefore take into account the fact that others will be trying to bring about the best outcome for themselves, and that they in turn will take into account the fact that I will be doing the same thing for myself.
- Finding the best action to perform is a kind of optimization problem, and optimization problems in general are hard to solve (typically NP-hard).

The first issue requires us to formulate a notion of rational outcome in a strategic setting. In game theory, such rational outcomes are given by *solution concepts*. A solution concept identifies a subset of the outcomes of a game—the rational outcomes according to the notion of rationality embodied within the solution concept. Typically, solution concepts capture some notion of equilibrium, or steady state—an outcome from which nobody has an incentive to deviate. As it turns out, there is no single good-for-all solution concept. Typical problems are that solution concepts sometimes fail to identify *any* outcomes of the game, and that they sometimes identify multiple outcomes. The second problem, where there are multiple outcomes, is a coordination problem, because the players in the game must coordinate on just one outcome. In the game theory literature, these are known as *equilibrium selection* problems.

Two Problematic Games

Two well-known games illustrate these concepts and highlight some of the problems that arise in game theoretic analysis. The first is probably the most famous in the game theory canon: the Prisoner's Dilemma. The Prisoner's Dilemma is usually introduced by way of the following story.

Alex and Bob are collectively charged with a crime and held in separate cells, with no way of meeting or communicating. They know that

- if one of them confesses to the crime and the other does not, the confessor will be freed, and the other will be jailed for three years (in England, we call this "turning Queen's evidence"—it tends to make you unpopular with other villains);
- if both confess, each will be jailed for two years; and
- if neither confesses, each will be jailed for one year.

The prisoners have to decide whether to cooperate (keep their mouths shut) or not cooperate (confess to the crime). How should Alex and Bob rationally decide between the two available choices? To answer this question, consider the following line of reasoning, from Alex's point of view:

- Suppose Bob confesses: if I confess, I would serve two years in prison,

Are We Rational?

A standard criticism of game theory is that solution concepts often fail to predict how people actually behave. One goal of behavioral economics is to try to understand how humans make decisions and to consider how such decision making relates to game-theoretic and other economic models. Dan Ariely recently popularized work in this area in his bestselling book *Predictably Irrational* (see the "Further Reading" sidebar). One short article cannot do justice to the whole of Ariely's hugely entertaining study of human irrationality, but this example gives the flavor of the kind of experiment he studies.

Ariely's aim was to study how "free" items affect human decision making. To investigate this, he offered students two types of chocolate for sale. One was a high-quality chocolate truffle, with a market value of 30 cents per piece; the other was a low-cost candy with a market value of only a couple of cents. When offered these chocolates at the price of 15 cents and 1 cent respectively, most students chose the truffle. They got a very good deal, effectively saving 15 cents on each truffle. This was the utility-maximizing choice: in both cases the students would benefit, but if they chose a truffle they would gain more. Then, Ariely reduced the price of both items by a cent, selling the truffle for 14 cents and giving the low-cost chocolate away for free. Presented with such an offer, students overwhelmingly chose the free item, despite the fact that the truffle clearly remained the utility-maximizing choice. The presence of a free item clearly seems to perturb our decision-making abilities away from the outcome that an economic/decision-theoretic analysis would suggest. Of course, this kind of irrational behavior is by no means limited to situations in which we are offered an item for free. Ariely investigates a range of situations in which human decision making seems to be skewed against rational outcomes.

		Alex	
		Confess	**Keep quiet**
Bob	**Confess**	–2, –2	0, –3
	Keep quiet	–3, 0	–1, –1

Figure 1. Payoff matrix for the Prisoner's Dilemma. The result pairs list Bob's result first, then Alex's.

and if I keep quiet, I would serve three years. Thus, my best choice in this case would be to confess.

- But suppose Bob keeps quiet: then, if I confess, I would go free, and if I keep quiet, I would spend a year in jail. My best choice in this case would also be to confess.

So, no matter what Bob does, Alex's best course of action is to confess.

Now, the Prisoner's Dilemma is symmetric: Bob will reason in the same way about Alex, and in the same way will conclude that no matter what choice Alex makes, his best choice would be to confess also. The upshot is that they both confess, and the overall outcome of the game is that Alex and Bob each serve two years in jail.

We can summarize the Prisoner's Dilemma using a standard piece of notation called a *payoff matrix*. Figure 1 shows the payoff matrix for the prisoner's dilemma:

Alex is the *column player*, so called because in making his choice, he chooses the column in which the outcome of the game will appear; Bob is the *row player*, who chooses the row in which the outcome of the game will appear. The four cells in the matrix represent the possible outcomes of the game. The values in the cells represent the *utility* obtained by each player from the corresponding outcome. Utility is simply a numeric measure of preferences: the idea is to attach a numeric value to outcomes, and we prefer outcomes with a higher utility value. In this case, the utilities correspond to the number of years in prison; the more years locked away, the worse the outcome for that player.

For example, the top right cell of the matrix is the outcome in which the column player, Alex, keeps quiet (cooperates), while the row player, Bob, confesses. This is the best possible outcome for Bob: he goes free, thereby obtaining a utility of 0 (no time in prison). However, it is the worst outcome for Alex (three years in jail, giving a utility of –3).

So, why is the Prisoner's Dilemma called a dilemma? Because if we consider the possibilities again, we can see that if both had kept quiet, the outcome would have been the cell in the bottom right of the payoff matrix, giving them each one year in prison—which is better for both of them than the mutual confession outcome. And the Prisoner's Dilemma is more troubling by the fact that it seems to correspond to many real-world scenarios.

For example, the tragedy of the commons seems to be a Prisoner's Dilemma. The tragedy of the commons occurs when there is some resource that can be shared by multiple agents but degrades rapidly if it is overused. One example is a piece of common land on which people can graze cattle. If everyone uses the common land carefully, the land will stay in good shape, but if it is overused, it becomes barren and is no good to anyone. The tragedy of the commons is a Prisoner's Dilemma because the best outcome for

Further Reading

D. Ariely, *Predictably Irrational*, Harper-Collins, 2008. A hugely entertaining study of human irrationality.

R. Axelrod, *The Evolution of Cooperation*, Basic Books, 1984. One of the most influential discussions on how cooperation can emerge from the actions of self-interested agents.

K, Binmore, *Game Theory: A Very Short Introduction*, Oxford Univ. Press, 2007. An enjoyable survey of game theory by one of its most prominent modern proponents.

J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton Univ. Press, 1944. Laid the foundations of contemporary game theory, although now perhaps of primarily historical interest.

N. Nisan et al., *Algorithmic Game Theory,* Cambridge Univ. Press, 2007. A collection of articles examining the role of game theory in theoretical computer science.

M. Osborne and A. Rubinstein, *A Course in Game Theory*, MIT Press, 1994. My favorite textbook on game theory.

W. Poundstone, *Prisoner's Dilemma*, Oxford Univ. Press, 1992. A popular history of game theory, with an emphasis on the Prisoner's Dilemma and its role in the analysis of cold war international relations.

Y. Shoham and K. Leyton-Brown, *Multiagent Systems: Algorithmic, Game-Theoretic, and Logical Foundations*, Cambridge Univ. Press, 2009. A good recent textbook on game theory from the viewpoint of multiagent systems and AI.

M. Wooldridge, *An Introduction to Multiagent Systems*, 2nd ed., John Wiley & Sons, 2009. An introduction to the scope and issues in multiagent systems research, mathematically lighter than the Shoham volume.

me is if I overuse the land but everyone else uses it carefully. But if we all act like this, we all overuse the land, and it becomes barren and no good for anybody. International fishing seems to be a rather sad contemporary example of the tragedy of the commons.

Nuclear arms reduction treaties are a (famous) example of the Prisoner's Dilemma. Suppose two countries, A and B, agree to eliminate their nuclear weapons. The best outcome for A would be that B gets rid of its weapons while A holds onto its weapons. But B will reason similarly, so both end up holding on to their weapons, despite the fact that they would both have been better off if they had gotten rid of them.

These outcomes, in which both players have an option that is the best response to all possible choices of the other player, are examples of a *dominant strategy equilibrium* solution concept. Dominant strategy equilibria do not always exist in games, but where they do, it is hard to see how any other outcome could rationally occur.

The Prisoner's Dilemma seems to be common, and the fact that the rational outcome, according to a game theoretic analysis, is worse for both players than another outcome has led many to claim that the game theoretic analysis must in some sense be wrong. (See the sidebar "Are We Rational?") A typical argument goes something like this: the game theoretic analysis says the only rational outcome is to confess rather than cooperate; but people do manage to cooperate in such situations, so the game theoretic analysis must be wrong.

Attempts to dismiss game theory on the basis of such analyses are often built on a misunderstanding of the scenario or what game theory has to say about it. In fact, there are quite natural variations of the Prisoner's Dilemma in which mutual cooperation can occur rationally—for example, if the participants play the game more than once. Intuitively, this is because players who do not "cooperate" can be "punished" in the future. However, we cannot possibly do justice to the substantial body of work here; Robert Axelrod, in *The Evolution of Cooperation*, provides one of the most influential discussions (see the "Further Reading" sidebar).

		Alex	
		Heads	**Tails**
Bob	**Heads**	1, –1	–1, 1
	Tails	–1, 1	1, –1

Figure 2. Payoff matrix for Matching Pennies. The result pairs list Bob's result first, then Alex's.

The second game, the game of Matching Pennies, is problematic for quite different reasons. In this case, Alex and Bob each have a coin, and each coin has two faces: heads and tails. They play a game in which each of them must simultaneously show just one side of his coin:

- If both coins show the same face, Bob wins: Alex pays Bob $1.
- If the coins show different faces, Alex wins: Bob pays Alex $1.

What should Alex and Bob do?

Figure 2 shows the payoff matrix for Matching Pennies.

There is no dominant strategy for Matching Pennies. If Bob shows heads, Alex's best response is to show tails, but if Bob shows tails, Alex's best response is to show heads. Similar reasoning applies for Bob's decision: neither choice is the best response to all choices of the other player.

So, is there another solution concept that we can apply here? Yes: the most famous solution concept of them all, known as Nash equilibrium. Intuitively, a Nash equilibrium is an outcome in which nobody wishes they had made a different choice, assuming that the other players stay with their choices. Nash equilibria thus capture the idea of a collection of individual choices being mutually the best response to each other. At first, it might appear that there is no Nash equilibrium in the game of Matching Pennies: if the coins are the same, Alex will regret his choice, whereas if they are different, Bob will regret his choice. However, the story is more subtle than this. Suppose both players randomize their choices; that is, play heads and tails with equal probability. This collection of strategies is, it turns out, a Nash equilibrium. In fact, John Nash demonstrated, in a famous result, that every game with a finite number of players and choices has a Nash equilibrium if we allow such randomization. (Technically, strategies in which we randomize in this way are called *mixed strategies*.) This is a powerful result indeed, and has led to Nash equilibria being regarded as the cornerstone of contemporary game theory research.

This brief article has introduced some of the main concepts of game theory—preferences, utilities, outcomes, solution concepts, dominant strategies, and Nash equilibria—and two of the best-known games in the game theory canon. These simple concepts are the foundations upon which contemporary game theory is built. Future articles will demonstrate how these concepts have been applied in AI, and also how AI has contributed to game theory itself. In the meantime, the "Further Reading" sidebar lists some of the most useful and influential books in the field.

Michael Wooldridge is a professor of computer science at the university of Liverpool. Contact him at mjw@liv.ac.uk.