Editor: **Michael Wooldridge,** University of Oxford, mjw@cs.ox.ac.uk

Does Game Theory Work?

Michael Wooldridge, *University of Oxford*

The past decade has witnessed a huge explosion of interest in issues that intersect computer science and game theory. We see game theory tracks at all major AI conferences, and algorithmic game theory has been one of the most high-profile growth areas in theoretical computer science in recent times. Given this level of interest, it might be worth stepping back and asking, "Does game theory actually work?" (with apologies to Ken Binmore, whose eponymous 2007 book inspired this article[1]). Here, I discuss some ways of viewing this question and review research on the topic.

It Depends What You Mean By "Work"

"Does game theory work?" seems at first sight to be a fairly straightforward question, but we need to be a little more precise before we can begin to answer it. Before we tighten up the question at hand, let's first recall what game theory is (and isn't).

Game theory is the mathematical theory of interactions between self-interested agents. In particular, it focuses on decision making in settings where each player's decision can influence the outcomes (and hence the well-being) of other players. In such settings, each player must consider how each other player will act in order to make an optimal choice. In game theory, the term "game" means an abstract mathematical model of a multiagent decision-making setting; the aim is typically to include in such a model all and only those aspects of the domain that are germane to the decisions that players must make. Game theory puts forward a number of *solution concepts* that are typically intended to formulate some notion of rational choice in a game-theoretic setting.

Solution concepts are thus at the heart of game theory. They're so called because they formulate solutions to games: game outcomes that could result if the players employed the corresponding notion of rational choice. *Nash equilibrium* is the most famous example of a solution concept. A Nash equilibrium is a selection of choices for players such that no player would prefer to unilaterally deviate from this selection.

So, given this discussion, how exactly are we to interpret game theory? Two obvious interpretations exist:

- Under a *descriptive interpretation*, we can view game theory as attempting to predict how (human) players will behave in strategic settings.
- Under a *normative interpretation*, we can view game theory as prescribing courses of action for players—that is, game theory tells players how they ought to act.

These two interpretations present very different criteria for the question of whether game theory works. The descriptive interpretation suggests that we should look for whether game theory successfully predicts how people will make choices in settings that we can model as games. The normative interpretation suggests that we should examine whether, by following game theory's prescriptions, we can obtain outcomes that are better than what we might otherwise have obtained. In the remainder of this article, I discuss these two interpretations and ask whether game theory "works" for them.

Descriptive Interpretations

Does game theory work as a tool for predicting how people will actually behave in game-like settings? Conventional wisdom says "no." However, the truth is more complex, subtle, and interesting than this simple answer suggests.

Social Norms

From as far back as the 1950s, when game theory was in its infancy, researchers wanted to investigate the extent to which the exciting new theorems and models they were developing could predict

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

actual behavior. Merrill Flood conducted one of the most famous early experiments to investigate this question.[2] This work introduced the prisoner's dilemma, now regarded as one of the most important formal games in the game theory canon. The point about the prisoner's dilemma is that the standard game-theoretic analysis leads to players selecting an outcome ("mutual defection") that's worse for both of them than another outcome ("mutual cooperation"). This fact has led many commentators to suggest that the game-theoretic analysis must be wrong.

Flood believed that "the axiomatic structures [of game theory] should be tested for applicability and usefulness in controlled experimental situations."[2] Jointly with Melvin Dresher, he organized a series of experiments (admittedly rather ad hoc experiments by today's rigorous standards) with the goal of doing just this. In one experiment, they had two people play the iterated prisoner's dilemma for 100 rounds; that is, they played a game with 100 rounds in which each round was the prisoner's dilemma. The players knew how long they were going to play for, and were both educated and mathematically proficient individuals with some knowledge of game theory. So, what did they do? Well, they didn't choose to follow the prescriptions of game theory, which in this case points to mutual defection in every round of the game. In fact, mutual cooperation occurred nearly two-thirds of the time, although the players' comments along the way ("dope … I'll be damned! ... the stinker! ... he's crazy … I'll teach him") suggest that the road to this outcome wasn't without some obstacles. Flood's interpretation was that the players were "rapidly learning … and converging to a split-the-difference [outcome]."[2]

Of course, we must treat Flood's experiments—and his analysis—with a good degree of caution. We can easily pick holes in the experimental setup and question the conclusions drawn. But this was among the first attempts to experimentally investigate whether game theory's solution concepts had any relation to actual observed human behavior, and one conclusion Flood drew—that the players were converging to "splitting the difference"—seems to reflect behavioral patterns that we see often when people interact.

We can better understand this behavior if we look at an even simpler experiment that Flood carried out,

> Social norms play an important role in determining how we behave when interacting with others.

which involved two secretaries playing the following game:

> The experimenter offers to give Subject 1 an amount $\$M$ but to give Subjects 1 and 2 together a greater amount $\$M + G$ if they can agree how to share the larger amount.

This is, of course, a rather cruel game. Subject 1 is clearly in a stronger position—he or she can always gain at least $\$M$, irrespective of Subject 2, and Subject 2 needs Subject 1's agreement to gain anything. Flood assumed that the secretaries would strike a deal in which Subject 1 received $\$M + (G/2)$, while Subject 2 received $\$G/2$—that is, the subjects would share the "surplus" G, while Subject 1 also received the entire amount M. What actually happened is that the secretaries divided the total amount $G + M$ equally between themselves, in fact resulting in Subject 1 receiving less than he or she would have simply by refusing to strike a deal with Subject 2. When questioned, the secretaries reported that they had agreed on this split in advance, but that if the values M and G in question were much larger, then they wouldn't have felt compelled to abide by the agreement. Flood's conclusion was that the social relationships between the secretaries acted to bring new factors into the game that weren't reflected in simply trying to maximize the portion of M and G that the subjects received. In other words, social factors change the utilities of the secretaries. Once again, we can easily criticize the experimental setup, but this misses the point, which is that even this crude experiment demonstrates that other factors are at work when players determine how to act. But what are those factors, exactly?

Subsequent research on this topic—and the literature is now enormous—suggests that social norms play an important role in determining how we behave when interacting with others. Social norms are nothing more than standards of behavior: conventions, or customs. They're rules of conduct that we adopt and follow as part of our upbringing and culture. We learn social norms from our parents ("Flush the toilet!"), from our friends at school ("Everyone should have a turn with the cool toy"), and when we start a job ("It isn't the done thing to hog the printer all the time"). Going back to our two secretaries, it seems likely that the social norm would simply

say that a windfall of this kind should be shared equally. As the stakes become higher, though, the temptation to transgress becomes greater. So what happens if one of the subjects transgresses—if Subject 1 insists on receiving $\$M + (G/2)$, for example? Nothing compels us to follow social norms, but we often do, because often such norms are self-policing. In the case of our two secretaries, the punishment for Subject 1 transgressing would be the shame of being viewed as greedy at a colleague's expense, or being forced to fetch your own coffee every day. But as the secretaries themselves indicated, for some sufficiently large value of $M + G$, Subject 1 would demand a larger piece of the pie: the benefits offered by $\$M + (G/2)$ start to outweigh the social stigma of greed or the tedium of always fetching your own coffee. Nevertheless, an enforcement mechanism appears to be behind the social norm here.

We generally view social norms—and, in particular, fairness norms such as splitting the difference—as playing havoc with game-theoretic models of rational choice. They seem to be responsible for the behavior in Flood's prisoner's dilemma experiments, and people seem to adhere to them quite commonly, both in real life and in experimental settings. So, where do they come from, and how and why do they persist?

Ken Binmore is a prominent game theorist with an interest in game theory's role in the social sciences. He's written extensively on game theory and social norms, and while I can't do justice to the breadth or depth of his viewpoints here, I can crudely summarize his views as follows:[3]

> A social norm (convention, social contract, and so on) is a self-policing agreement between members of a society that allows them to coordinate on a particular outcome.

The outcome in question will often be one that's efficient; such norms can potentially explain why people actually manage to cooperate (as opposed to defecting) in the prisoner's dilemma.

Models of Utility

There are other reasons why decisions people make in life and in the laboratories of game theorists don't necessarily correspond to the solution concepts we find in textbooks. One important one is that the utility model used in game-theoretic analysis doesn't necessarily correspond to how people assess utility. Perhaps the most famous work in this area comes from Daniel Kahneman and Amos Tversky—work for which Kahneman received a Nobel prize in 2002 (Tversky passed away in 1996). They put forward a theory of how people make decisions under uncertainty called *prospect theory*.[4] Their starting point was the observation that expected utility theory doesn't predict how people make decisions in practice. Expected utility theory is the best-known and most widely studied theory of decision making under uncertainty. The mathematics of expected utility theory is elegant and compelling, but a wealth of evidence shows that it doesn't fare well at predicting how people make decisions. This is easy to demonstrate. Consider the following game:[4]

> People tend to be loss-averse: they will make choices to avoid potential perceived losses.

You are given a choice between options A and B.

A: We flip a fair coin. If it comes up heads, you get $1,000; if it comes up tails, you get nothing.

B: You get $450.

Would you choose A or B? Expected utility theory says that you can expect to earn $500 from option A or $450 from option B, so option A is the rational choice. But I wouldn't choose A, nor, I think, would most people; I know how grumpy I would be if the coin came up tails, and it isn't worth the chance of that extra $50. (If we were to play the game 100 times, I would choose differently.) Even with such a simple experiment, expected utility theory fails.

Kahneman and Tversky's prospect theory suggests that when making decisions, people will broadly behave as follows. First, they don't dispassionately view all outcomes on some simple linear scale, with higher being better and lower being worse. Rather, they evaluate outcomes with respect to some fixed reference point, which captures a status quo. People view outcomes below this reference point as losses, and outcomes above this point as gains. In the aforementioned example, you might fix on $450 as a reference point, because you're guaranteed this amount by choosing option B. Alternatively, consider somebody who receives a bonus at work. Suppose they believed in advance that they would receive $50,000; this becomes the reference point. If they then in fact receive $100,000, they perceived this as a gain—it's above the reference point. In contrast, if

they receive $40,000, then they perceive this as a loss, despite the fact that they're considerably better off than they were originally. Prospect theory suggests that people tend to be loss-averse: they will make choices to avoid potential perceived losses.

Prospect theory is by no means the only alternative to expected utility theory that's been proposed, although it's probably the most successful such approach in recent times. Irrespective of prospect theory's advantages and disadvantages as a tool for predicting human decisions, however, we can see that conventional game theory assumes a model of decision making that doesn't necessarily reflect how people actually make choices.

Lessons

So what can we learn from the previous discussion?[1]

- In real-life settings, social norms (and, in particular, norms of cooperation) often play a part in how people make decisions. However, if the incentives at hand are sufficiently large, then these incentives can start to override social norms.
- More generally, for incentives (such as payments) to be sufficient to influence behavior, they must be adequate. Many game theory laboratory experiments fail because the rewards involved are insufficient to influence behavior.
- The model of choice that people use will not always be based on expected utility-type models. We must understand their baseline expectations and whether they see outcomes as gains or losses; and we must understand the tendency to be loss-averse.
- For players to make rational choices, the game they're playing must be sufficiently simple.
- Players will adapt their behavior over time toward more rational outcomes, if they receive sufficient opportunity for trial-and-error learning.

Much work remains on reconciling the theoretical models and solutions of game theory with observed human behavior. Such issues are currently studied in the field of behavioral game theory.

Normative Interpretations

Let's now turn to the normative interpretation of game theory, which describes what players ought to do in a game. Under a normative interpretation, we think of game theory as providing advice to players about how best to play games, or to game designers about how best to design them (this latter issue is called *mechanism design*). The normative interpretation works if the advice is good—that is, if it helps people to make better decisions or helps governments or other designers of economic mechanisms to design better ones. So, what evidence is there that game-theoretic advice is good?

> Often, game-theoretic models are applied in entirely inappropriate circumstances.

The first point to make here is that often, game-theoretic models are applied in entirely inappropriate circumstances. Such models are predicated on a host of assumptions—some that are easily justifiable, others that are perhaps harder to justify. To pick just one example, a standard game-theoretic assumption is that players have common knowledge of the game at hand: everybody knows who the players are, their available choices, and their preferences, and everybody knows that everybody knows this, and so on. It isn't hard to see that such assumptions simply don't hold in many settings, and where they don't hold, we can't hope to rely on game-theoretic advice. So, if we want to use a game-theoretic model to obtain advice about how to act, we had better first be sure that the model and its assumptions really fit the circumstances at hand.

Despite this word of caution, some prominent examples exist of game-theoretic advice being successfully used in important real-world settings. Let's briefly look at two.

Auction Design

In the 1990s, cell phones became widely used, and governments worldwide acted to regulate access to the electromagnetic frequencies that these phones use to communicate. At the end of the 1990s, mobile phone technology was moving toward a "third-generation" (3G) set of standards, enabling phones to access rich data services. These 3G services used a new range of frequencies, and many governments decided to use auctions to allocate these frequencies to 3G service providers. They recruited game theorists to design these auctions. In the UK, a team of game theorists was charged with designing auctions that would allocate frequency ranges with the following goals:[5]

- assign the electromagnetic spectrum efficiently—that is, allocate spectrum to those who would best use it;
- promote competition; and
- realize the spectrum's full economic value—that is, maximize revenue.

Much effort went into designing the auctions that were ultimately used, and the design process included laboratory experiments to determine how individuals would actually behave for various sets of possible auction scenarios. The auctions themselves occurred in the first half of 2000 (which was, coincidentally, the peak of the dot-com boom, when wild speculative investment was made in IT and telecommunications technology). The results were staggering. In total, the auctions raised $34 billion in revenue for the UK government. This is an astonishing amount: to recoup their investment, those that obtained the licenses would have to receive payment for 3G services of at least $550 for every man, woman, and child in the UK! Although the UK government was naturally jubilant at this windfall, telecom companies quickly regretted the scale of their ambitions and began to complain about how much they had paid for licenses. Whatever side you take in this debate (and I have to say, complaints about the auctions weren't received with much sympathy by the game theorists who designed them), the game-theoretic auction design process certainly paid off spectacularly from the government's viewpoint.

Security Games

Since 9/11, the world has become security obsessed. Time-consuming security checks and screening procedures are now standard in international travel, and those charged with ensuring our personal safety invest huge amounts of time and money into these checks and procedures. But despite their best efforts, security is an inexact science. One key problem is that the resources available to security organizations are inherently limited, in which case a fundamental problem is how to allocate these scarce resources to best effect. This problem has provided game theory with one of its most innovative and compelling recent application areas. The basic idea is to view the problem as a game played between the security organization and its adversaries. The moves available to the adversary typically correspond to attacking one of the defender's assets, while the moves available to the defender typically correspond to allocating security teams to these assets. The game-theoretic analysis provides a randomized strategy, which indicates how to assign security resources to assets. These techniques have been deployed in numerous real-world settings. Most famously, they currently inform security strategy at Los Angeles World Airports (LAX). The security games paradigm was largely the work of Milind Tambe and his group, and Tambe's book[6] represents the state of the art in this most challenging domain.

> Game theory can work under both its descriptive and normative interpretations, although it might often appear that it doesn't.

As is often the case in science, the answer to an apparently simple question has a frustratingly complex answer, wrapped almost to obscurity in caveats and disclaimers. Game theory can work under both its descriptive and normative interpretations, although it might often appear that it doesn't. Applied in the right way—by understanding the motivations of those who participate and the social norms that guide them—game theory can successfully predict how people will behave and, similarly, its techniques can help in economic system design. But of course, much research remains before we can fully understand the scope, applicability, and usefulness of game-theoretic techniques. Expect to see much more work on this subject in the years to come.

References

1. K. Binmore, *Does Game Theory Work? The Bargaining Challenge*, MIT Press, 2007.
2. M.M. Flood, "Some Experimental Games," research memorandum RM-789, RAND Corp., 1952; www.rand.org/pubs/.
3. K. Binmore, *Playing Fair*, MIT Press, 1994.
4. D. Kahneman and A. Tversky, "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, vol. 47, no. 2, 1979, pp. 263–292.
5. P. Klemperer and K. Binmore, "The Biggest Auction Ever: The Sale of the British 3G Telecom Licenses," working paper, University College London, 2001.
6. M. Tambe, *Security Games*, Cambridge Univ. Press, 2012.

Michael Wooldridge is a professor of computer science at the University of Oxford. Contact him at mjw@cs.ox.ac.uk.