AMERICAN
Scientist

BOOK REVIEW

Honor Among Thieves

Cosma Shalizi

THE CALCULUS OF SELFISHNESS. Karl Sigmund. x + 173 pp. Princeton University Press, 2010. $35.

Since the 1970s, a loose community of theoretical biologists, economists, political scientists, mathematicians and philosophers has been using the tools of evolutionary game theory to try to understand how purely selfish agents can come to cooperate, follow norms and even behave altruistically—to understand when honesty is the best policy. Karl Sigmund has been a leading figure in these efforts, and *The Calculus of Selfishness* is his latest attempt at an introduction to the field. In its exposition, the book focuses on reciprocity between self-interested individuals in certain elementary types of interactions.

Game theory (as laid out in chapter 1) models agents interacting with each other, in pairs or in larger groups, with a fixed set of moves available to them. (Sigmund mostly deals with two-player games, but this is just for simplicity.) An agent—Alice, say—has a "strategy" that tells her what move to make at each step in the game, in response to another agent's moves and to the state of the external world (if the model admits that the latter exists). Alice's coplayer—Bob, say—also has a strategy, and together the two strategies determine the outcome of the game. At its end, Alice and Bob each get a *payoff,* according to a function that depends on the moves both have made. Their strategies are in equilibrium if neither of them could increase their payoff by changing their moves unilaterally.

Basic economics courses lead one to expect that there will be only one equilibrium and that it will be optimal for everyone. The games of relevance to the evolution of cooperation, however, are *social dilemmas,* where this expectation fails. Sometimes the problem is that the equilibrium is optimal for no one. Imagine that Alice and Bob are two bandits, who can either *cooperate* in robbing villages and caravans, or *defect* by turning on each other. If they both cooperate, each will take $1,000; if they both defect, neither can steal effectively and they'll get $0. But suppose that if Alice cooperates and Bob defects by turning on her, he will get $2,000 and she will lose $500—and vice versa. Then regardless of what Alice does, Bob will be better off defecting. So by symmetry, the only way to achieve equilibrium is for both of them to defect—even though they'd both be better off, in purely selfish terms, if they both cooperated. When this type of problem was first posed at the RAND Corporation in the early 1950s, it was framed as prisoners being offered the chance to turn state's witness, so it is still called the *prisoner's dilemma,* and it is the archetype of many cooperation problems. A multiperson version of it is the *public goods* game, in which something like an open park is shared equally among all participants, no matter how much each of them has contributed to its creation or upkeep.

In other social dilemmas, there are many equilibria, some more favorable to one player than to another. If Alice and Bob can agree to cooperate as bandits, how should they split the loot—the extra loot, that is, beyond what they would each get on their own? If Alice can convince Bob that she won't cooperate unless she gets, say, 75 percent of the gains, it is in Bob's interest to agree to her demand, since 25 percent extra is better than 0 percent. Thus there is an Alice-favoring equilibrium; but, symmetrically, there is also a Bob-favoring one. However, agreeing to a 50-50 split is not an equilibrium at all. This is called the *snowdrift game,* after another fable, and is the archetypal depiction of the inevitable struggle over how to divide the gains from cooperation.

In evolutionary game theory (introduced in chapter 2), we imagine a population of agents, all with their own strategies, who are randomly paired up to play a given game. The *replicator dynamic* means that strategies whose players do better than average will increase their share of the population, at the expense of strategies that have worse-than-average results. Whether this works by agents imitating the more successful among them, or by actual natural selection, is not (at this level of abstraction) mathematically relevant. What matters is that the performance of each strategy depends on what other strategies are available in the population and how common they are. The population-level dynamics of strategies are stochastic processes, which in the large-population limit (where Sigmund mostly works) smooth out into differential equations.

When are cooperative strategies able to survive or even dominate the population? Chapters 3, 4 and 5 consider mechanisms that can achieve this. Chapter 3 is about direct reciprocity: If Alice and Bob play the same game with each other repeatedly, and each can remember what the other player did earlier, they can pay each other back in kind. In games of the prisoner's dilemma type, a strategy of cooperating if and only if the other player cooperated last time does not do well against a fixed strategy of noncooperation, but it *does* do well against another player with the same strategy. It can thus invade a population of noncooperators, although once it becomes dominant, a strategy of unconditional cooperation can invade in turn. These strategies require only very basic sorts of memory; more complex strategies, relying on more sophisticated memories, can be more robust to noise, avoiding costly cycles of retaliation provoked by error and misunderstanding.

Chapter 4 extends these ideas to *indirect reciprocity,* in two forms. In both, Alice's decision to cooperate with Bob today leads to cooperation tomorrow. In one form, today's cooperation leads Charlie to cooperate with Alice tomorrow; in the other, it leads Bob to cooperate with Charlie.

Chapter 5 looks at norms of fairness and the effects of having a reputation for refusing to violate norms. In the split-the-loot game above, if Alice has established a reputation for refusing to cooperate at all rather than accept unfair divisions, Bob will not bother to offer her a bad deal in the first place. Both indirect reciprocity and reputation require not just memory but also fairly accurate information about what third parties have done—information acquired by perception or communication.

Chapter 6 looks in detail at a particular model of the provision of public goods, and at whether voluntary or compulsory participation in providing public goods works better. This chapter is exemplary as an analysis of a model, but the conclusions seem to rest heavily on the assumption that each agent's reward from the public good goes down in proportion to the population size, which is not true of many public goods, such as roads free of bandits.

Up to this point, the book has assumed that players meet randomly, which is of course unrealistic. The final chapter looks briefly at what happens when some internal structure is added to the populations, by spreading them out in space and limiting agents to playing with their immediate neighbors, or by dividing the population into subgroups that mix internally and periodically re-form. This chapter feels less complete than the rest, because it offers just a glimpse of a vast, largely unmapped territory.

Sigmund's mathematical exposition is exemplary. He starts with the presumption that the reader has only rudimentary linear algebra (multiplying vectors by matrices) and some notion of what a differential equation is, and he builds up from there, introducing more advanced concepts and results as needed. He avoids formal proofs and bookkeeping in favor of careful explanations of key points and illustrative calculations. As he teaches evolutionary game theory, Sigmund is also demonstrating how to write about applied mathematics.