PRINCETON SERIES IN THEORETICAL AND COMPUTATIONAL BIOLOGY

Series Editor, Simon A. Levin

The Calculus of Selfishness,
by Karl Sigmund

The Geographic Spread of Infectious Diseases: Models and Applications,
by Lisa Sattenspiel with contributions from Alun Lloyd

Theories of Population Variation in Genes and Genomes,
by Freddy Bugge Christiansen

Analysis of Evolutionary Processes,
by Fabio Dercole and Sergio Rinaldi

Mathematics in Population Biology,
by Horst R. Thieme

Individual-based Modeling and Ecology,
by Volker Grimm and Steven F. Railsback

The Calculus of Selfishness

Karl Sigmund

PRINCETON UNIVERSITY PRESS
PRINCETON AND OXFORD

Chapter One

Introduction: Social Traps and Simple Games

1.1 THE SOCIAL ANIMAL

Aristotle classified humans as social animals, along with other species, such as ants and bees. Since then, countless authors have compared cities or states with bee hives and ant hills: for instance, Bernard de Mandeville, who published his *The Fable of the Bees* more than three hundred years ago.

Today, we know that the parallels between human communities and insect states do not reach very far. The amazing degree of cooperation found among social insects is essentially due to the strong family ties within ant hills or bee hives. Humans, by contrast, often collaborate with non-related partners.

Cooperation among close relatives is explained by *kin selection*. Genes for helping offspring are obviously favoring their own transmission. Genes for helping brothers and sisters can also favor their own transmission, not through direct descendants, but indirectly, through the siblings' descendants: indeed, close relatives are highly likely to also carry these genes. In a bee hive, all workers are sisters and the queen is their mother. It may happen that the queen had several mates, and then the average relatedness is reduced; the theory of kin selection has its share of complex and controversial issues. But family ties go a long way to explain collaboration.

The bee-hive can be viewed as a watered-down version of a multicellular organism. All the body cells of such an organism carry the same genes, but the body cells do not reproduce directly, any more than the sterile worker-bees do. The body cells collaborate to transmit copies of their genes through the germ cells—the eggs and sperm of their organism.

Viewing human societies as multi-cellular organisms working to one purpose is misleading. Most humans tend to reproduce themselves. Plenty of collaboration takes place between non-relatives. And while we certainly have been selected for living in groups (our ancestors may have done so for thirty million years), our actions are not as coordinated as those of liver cells, nor as hard-wired as those of social insects. Human cooperation is frequently based on individual decisions guided by personal interests.

Our communities are no super-organisms. Former Prime Minister Margaret Thatcher pithily claimed that "there is no such thing as society." This can serve as the rallying cry of *methodological individualism*—a research program aiming to explain collective phenomena bottom-up, by the interactions of the individuals involved. The mathematical tool for this program is game theory. All "players" have their own aims. The resulting outcome can be vastly different from any of these aims, of course.

1.2 THE INVISIBLE HAND

If the end result depends on the decisions of several, possibly many individuals having distinct, possibly opposite interests, then all seems set to produce a cacophony of conflicts. In his *Leviathan* from 1651, Hobbes claimed that selfish urgings lead to "such a war as is every man against every man." In the absence of a central authority suppressing these conflicts, human life is "solitary, poore, nasty, brutish, and short." His French contemporary Pascal held an equally pessimistic view: "We are born unfair; for everyone inclines towards himself. . . . The tendency towards oneself is the origin of every disorder in war, polity, economy etc." Selfishness was depicted as the root of all evil.

But one century later, Adam Smith offered another view. An invisible hand harmonizes the selfish efforts of individuals: by striving to maximize their own revenue, they maximize the total good. The selfish person works inadvertently for the public benefit. "By pursuing his own interest he frequently promotes that of the society more effectually than when he really intends to promote it." Greed promotes behavior beneficial to others. "It is not from the benevolence of the butcher, the brewer, or the baker, that we expect our dinner, but from their regard to their own self-interest. We address ourselves, not to their humanity but to their self-love, and never talk to them of our own necessities but of their advantages."

A similar view had been expressed, well before Adam Smith, by Voltaire in his *Lettres philosophiques*: "Assuredly, God could have created beings uniquely interested in the welfare of others. In that case, traders would have been to India by charity, and the mason would saw stones to please his neighbor. But God designed things otherwise. . . . It is through our mutual needs that we are useful to the human species; this is the grounding of every trade; it is the eternal link between men."

Adam Smith (who knew Voltaire well) was not blind to the fact that the invisible hand is not always at work. He merely claimed that it *frequently* promotes the interest of the society, not that it always does. Today, we know that there are many situations—so-called social dilemmas—where the invisible hand fails to turn self-interest to everyone's advantage.

1.3 THE PRISONER'S DILEMMA

Suppose that two individuals are asked, independently, whether they wish to give a donation to the other or not. The donor would have to pay 5 dollars for the beneficiary to receive 15 dollars. It is clear that if both players cooperate by giving a donation to their partner, they win 10 dollars each. But it is equally clear that for each of the two players, the most profitable strategy is to donate nothing, i.e., to defect. No matter whether your co-player cooperates or defects, it is not in your interest to part with 5 dollars. If the co-player cooperates, you have the choice between obtaining, as payoff, either 15 dollars, or 10. Clearly, you should defect. And if the co-player defects, you have the choice between getting nothing, or losing 5 dollars. Again, you should defect. To describe the Donation game in a nutshell:

		if the co-player makes a donation	if the co-player makes no donation
My payoff	if I make a donation	10 dollars	−5 dollars
	if I make no donation	15 dollars	0 dollars

But the other player is in the same situation. Hence, by pursuing their selfish interests, the two players will defect, producing an outcome that is bad for both. Where is the invisible hand? "It is often invisible because it is not here," according to economist Joseph Stiglitz.

This strange game is an example of a *Prisoner's Dilemma*. This is an interaction between two players, player I and II, each having two options: to cooperate (play C) or to defect (play D). If both cooperate, each obtains a *Reward* R that is higher than the *Punishment* P, which they obtain if both defect. But if one player defects and the other cooperates, then the defector obtains a payoff T (the *Temptation*) that is even higher than the Reward, and the cooperator is left with a payoff S (the *Sucker's payoff*), which is lowest of all. Thus,

$$T > R > P > S. \tag{1.1}$$

As before, it is best to play D, no matter what the co-player is doing.

		if player II plays C	if player II plays D
Payoff for player I	if player I plays C	R	S
	if player I plays D	T	P

If both players aim at maximizing their own payoff, they end up with a suboptimal outcome. This outcome is a trap: indeed, no player has an incentive to switch unilaterally from D to C. It would be good, of course, if both *jointly* adopted C. But as soon as you know that the other player will play C, you are faced with the Temptation to improve your lot still more by playing D. We are back at the beginning. The only consistent solution is to defect, which leads to an economic stalemate.

The term "Prisoner's Dilemma" is used for this type of interaction because when it was first formulated, back in the early fifties of last century, it was presented as the story of two prisoners accused of a joint crime. In order to get confessions, the state attorney separates them, and proposes a deal to each: they can go free (as state's witness) if they rat on their accomplice. The accomplice would then have to face ten years in jail. But it is understood that the two prisoners cannot *both* become state's witnesses: if both confess, both will serve seven years. If both keep mum, the attorney will keep them in jail for one year, pending trial. This is the original Prisoner's Dilemma. The Temptation is to turn state's witness, the Reward consists in being released after the trial, (which may take place only one year from now), the Punishment is the seven years in jail and the Sucker's payoff amounts to ten years of confinement.

The young mathematicians who first investigated this game were employees of the Rand Corporation, which was a major think tank during the Cold War. They may have been inspired by the dilemma facing the two superpowers. Both the Soviet Union and the United States would have been better off with joint nuclear disarmament. But the temptation was to keep a few atomic bombs and wait for the others to destroy their nuclear arsenal. The outcome was a horrendously expensive arms race.

1.4 THE SNOWDRIFT GAME

The Prisoner's Dilemma is not the only social dilemma displaying the pitfalls of selfishness. Another is the so-called *Snowdrift* game. Imagine that the experimenter promises to give the two players 40 dollars each, on receiving from them a "fee" of 30 dollars. The two players have to decide separately whether they want to come up with the fee, knowing that if they both do, they can share the cost. This seems to be the obvious solution: they would then invest 15 dollars each, receive 40 in return, and thus earn 25 dollars. But suppose that one player absolutely refuses to pay. In that case, the other player is well advised to come up with 30 dollars, because this still leads to a gain of 10 dollars in the end. The decision is hard to swallow, however, because the player who invests nothing receives 40 dollars. If both players are unwilling to pay the fee, both receive nothing. This can be described

		if my co-player contributes	if my co-player refuses to contribute
My payoff	if I contribute	25, 25	10, 40
	if I refuse to contribute	40, 10	0, 0

as a game with the two options C (meaning to be willing to come up with the fee) and D (not to be willing to do so). If we denote the payoff values with R,S,T, and P, as before, we see that in the place of (equation 1.1.) we now have

$$T > R > S > P. \tag{1.2}$$

Due to the small difference in the rank-ordering (only S and P have changed place), playing D is not *always* the best move, irrespective of the co-player's decision. If the co-player opts for D, it is better to play C. In fact, for both players, the best move is to do the opposite of what the co-player decides. But in addition, both know that they will be better off by being the one who plays D. This leads to a contest. If both insist on their best option, both end up with the worst outcome. One of them has to yield. This far the two players agree, but that is where the agreement ends.

The name *Snowdrift* game refers to the situation of two drivers caught with their cars in a snow drift. If they want to get home, they have to clear a path. The fairest solution would be for both of them to start shoveling (we assume that both have a shovel in their trunk). But suppose that one of them stubbornly refuses to dig. The other driver could do the same, but this would mean sitting through a cold night. It is better to shovel a path clear, even if the shirker can profit from it without lifting a finger.

1.5 THE REPEATED PRISONER'S DILEMMA

The prisoners, the superpowers, or the test persons from the economic experiments may seem remote from everyday life, but during the course of a day, most of us will experience several similar situations in small-scale economic interactions. Even in the days before markets and money, humans were engaged in ceaseless give and take within their family, their group or their neighborhood, and faced with the temptation to give less and take more.

The artificial aspect of the Donation game is not due to its payoff structure, but to the underlying assumption that the two players interact just once, and then go their separate ways. Most of our interactions are with household members, colleagues, and other people we are seeing again and again.

The games studied so far were *one-shot* games. Let us assume now that the same two players repeat the same game for several rounds. It seems obvious that a player who yields to the temptation of exploiting the co-player must expect retaliation. Your move in one round is likely to affect your co-player's behavior in the following rounds.

Thus let us assume that the players are engaged in a Donation game repeated for six rounds. Will this improve the odds for cooperation? Not really, according to an argument called *backward induction*. Indeed, consider the sixth and last round of the new game. Since there are no follow-up rounds, and since what's past is past, this round can be viewed in isolation. It thus reduces to a one-shot Donation game, for which selfish interests, as we have seen, prescribe mutual defection. This is the so-called "last-round effect." Both players are likely to understand that nothing they do can alter this outcome. Hence, they may just as well take it for granted, omit it from further consideration, and just deal with the five rounds preceding the last one. But for the fifth round, the same argument as before prescribes the same move, leading to mutual defection; and so on. Hence backward induction shows that the players should never cooperate. The players are faced with a money pump that can deliver 10 dollars in each round, and yet their selfish interests prescribe them not to use it. This is bizarre. It seems clearly smarter to play C in the first round, and signal to the co-player that you do not buy the relentless logic of backward induction.

It is actually a side-issue. Indeed, people engaged in ongoing everyday interactions do rarely know beforehand which is the last round. Usually, there is a possibility for a further interaction—the *shadow of the future*. Suppose for instance that players are told that the experimenter, after each round, throws dice. If it is six, the game is stopped. If not, there is a further round of the Donation game, to be followed again by a toss of the dice, etc. The duration of the game, then, is random. It could be over after the next round, or it could go on for another twenty rounds. On average, the game lasts for six rounds. But it is never possible to exploit the co-player without fearing retaliation.

In contrast to the one-shot Prisoner's Dilemma, there now exists no strategy that is best against all comers.If your co-player uses an unconditional strategy and always defects, or always cooperates, come what may, then it is obviously best to always defect. However, against a touchy adversary who plays C as long as you do, but turns to relentlessly playing D after having experienced the first defection, it is better to play C in every round. Indeed, if you play D, you exploit such a player and gain an extra 5 dollars; but you lose all prospects of future rewards, and will never obtain a positive payoff in a further round. Since you can expect that the game lasts for 5 more rounds, on average, you give up 50 dollars.

What about the repeated Snowdrift game? It is easy to see that if the two players both play C in each round, or if they alternate in paying the fee, i.e., being the C player, then they will both do equally well, on average; but is it likely that they will reach such a symmetric solution? Should we rather expect that one of the two players gives in, after a few rounds, and accepts grudgingly the role of the exploited C player? The joint income, in that case, is as good as if they both always cooperate, but the distribution of the income is highly skewed.

1.6 TOURNAMENTS

Which strategy should you chose for the repeated Prisoner's Dilemma, knowing that none is best? Some thirty years ago, political scientist Robert Axelrod held a computer tournament to find out. People could submit strategies. These were then matched against each other, in a round-robin tournament: each one engaged each other in an iterated Prisoner's Dilemma game lasting for 200 rounds (the duration was not known in advance to the participants, so as to offer no scope for backward induction). Some of the strategies were truly sophisticated, testing out the responses of the co-players and attempting to exploit their weaknesses. But the clear winner was the simplest of all strategies submitted, namely *Tit for Tat* (*TFT*), the epitome of all retaliatory strategies. A player using *TFT* plays C in the first move, and from then on simply repeats the move used by the co-player in the previous round.

The triumph of *TFT* came as a surprise to many. It seemed almost paradoxical, since *TFT* players can *never* do better than their co-players in a repeated Prisoner's Dilemma game. Indeed, during the sequence of rounds, a *TFT* player is never ahead. As long as both players cooperate, they do equally well. A co-player using D draws ahead, gaining T versus the *TFT* player's payoff S. But in the following rounds, the *TFT* player loses no more ground. As long as the co-player keeps playing D, both players earn the same amount, namely P. If the co-player switches back to C, the *TFT* player draws level again, but resumes cooperation forthwith. At any stage of the game, *TFT* players have either accumulated the same payoff as their adversary, or are lagging behind by the payoff difference $T - S$. But in Axelrod's tournament, the payoffs against all co-players had to be added to yield the total score; and thus *TFT* ended ahead of the rest, by doing better than every co-player *against the other entrants*.

Axelrod found that among the 16 entrants for the tournaments, eight were *nice* in the sense that they never defected first. And these eight took the first eight places in

the tournament. Nice guys finish first! In fact, Axelrod found that another strategy even "nicer" than *TFT* would have won the tournament, had it been entered. This was *TFTT* (*Tit for Two Tats*), a strategy prescribing to defect only after two consecutive D's of the co-player. When Axelrod repeated his tournaments, 64 entrants showed up, and one of them duly submitted *TFTT*. But this strategy, which would have won the first tournament, only reached rank 21. Amazingly, the winner of the second tournament was again the simplistic *TFT*. It was not just nice, it was transparent, provokable, forgiving, and robust.This bouquet of qualities seemed the key to success.

1.7 ARTIFICIAL SOCIETIES

The success of Axelrod's tournaments launched a flurry of computer simulations. Individual-based modeling of artificial societies greatly expanded the scope of game theory. Artificial societies consist of fictitious individuals, each equipped with a strategy specified by a program. These individuals meet randomly, engage in an iterated Prisoner's Dilemma game, and then move on to meet others. At the end, the accumulated payoffs are compared. Often, such a tournament is used to update the artificial population. This means that individuals produce "offspring", i.e., other fictitious individuals inheriting their strategy. Those with higher payoffs produce more individuals, so that successful strategies spread. Alternatively, instead of inheriting strategies, the new individuals can adapt by copying strategies, preferentially from individuals who did better. In such individual-based simulations, the frequencies of the strategies change with time. One can also occasionally introduce small minorities using new strategies, and see whether these spread or not. In chapter 2, we shall describe the mathematical background to analyze such models.

Let us consider, for instance, a population using only two strategies, *TFT* and *AllD*. The average payoff for a *TFT* player against another is 60 dollars (corresponding to 6 rounds of mutual cooperation). If a *TFT* player meets an *AllD* player, the latter obtains 15 dollars (by exploiting the co-player in the first round) and the former loses 5 dollars. If two *AllD* players meet each other, they get nothing.

		if the co-player plays *Tit for Tat*	if the co-player always defects
My payoff	if I play *Tit for Tat* (*TFT*)	60	−5
	if I always defect (*AllD*)	15	0

Players having to choose among these two strategies fare best by doing what the co-player does, i.e., playing *TFT* against a *TFT* player and *AllD* against an *AllD* player. But in individual-based modeling, the fictitious players have no options. They are stuck with their strategy, and do not know their co-player's strategy in advance. Obviously, the expected payoff depends on the composition of the artificial population. If most play *TFT*, then *TFT* is favored; but in a world of defectors, *AllD* does better. In the latter case, the players are caught in a social trap. Games with