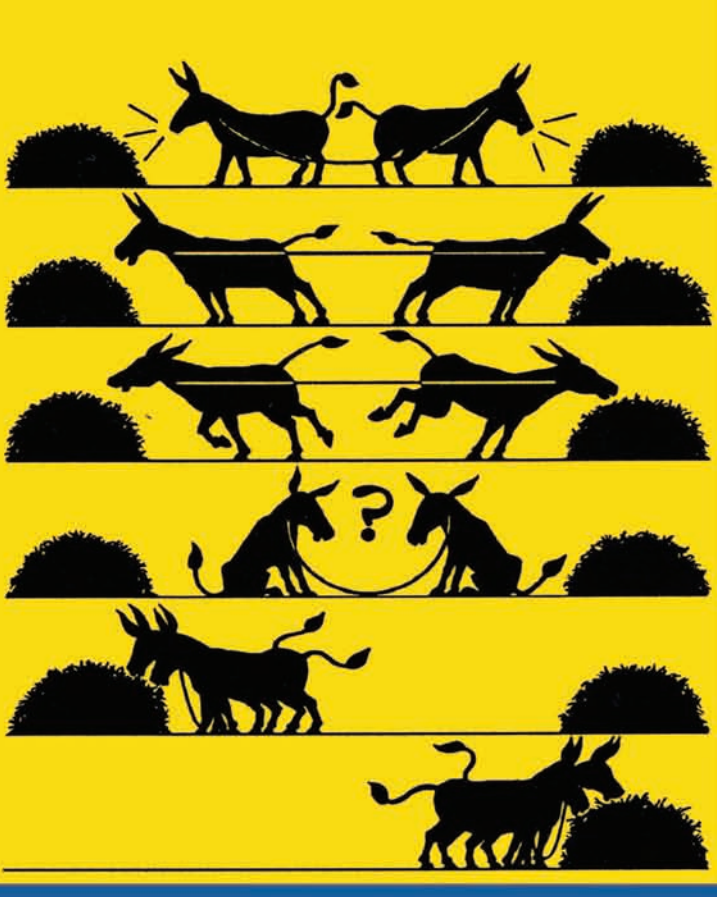

PRINCETON SERIES IN THEORETICAL AND COMPUTATIONAL BIOLOGY

The Calculus of Selfishness

KARL SIGMUND

The Calculus of Selfishness

PRINCETON SERIES IN THEORETICAL AND COMPUTATIONAL BIOLOGY

Series Editor, Simon A. Levin

The Calculus of Selfishnes ,
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

Published by Princeton University Press, 41 William Street, Princeton, New Jersey 08540

In the United Kingdom: Princeton University Press, 6 Oxford Street, Woodstock,
Oxfordshire OX20 1TW

Library of Congress Cataloging-in-Publication Data

Sigmund, Karl, 1945-
The calculus of selfishness / Karl Sigmund.
p. cm. — (Princeton series in theoretical and computational biology)
Includes bibliographical references and index.
ISBN 978-0-691-14275-3 (hardcover : alk. paper) 1. Game theory. 2. Cooperativeness—Moral and ethical aspects. 3. Evolution (Biology)—Mathematics. I. Title.
HB144.S59 2009
306.3′4—dc22 2009015030

British Library Cataloging-in-Publication Data is available

This book has been composed in Times & Abadi

Printed on acid-free paper. ∞

press.princeton.edu

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

Contents

Preface

You need not be a scheming egotist to pick up *The Calculus of Selfishnes* . It is enough to be interested in the logic behind the ceaseless give-and-take pervading our social lives. The readership I had in mind, when writing this book, consists mostly of undergraduates in economics, psychology, or evolutionary biology. But simple models of social dilemmas are of general interest.

As the word *Calculus* in the title gives away, you will need a modicum of elementary mathematics. Beyond this, all the game-theory expertise you need is painlessly provided step by step. As to the *Selfishnes* in the title, I do not mean blind greed, of course, but "enlightened self-interest," by which, according to Tocqueville, "Americans are fond of explaining almost all the actions of their lives; . . . . They show with complacency how an enlightened regard for themselves constantly prompts them to assist each other." Such complacency may well be justified; but theoreticians cannot share it. Most of them feel that it is hard to understand why self-interested agents cooperate for their common good.

In the New Year 2000 edition of *Science*, the editors listed "The evolution of cooperation" as one of the ten most challenging problems of the century. My book certainly does not claim to solve the problem. Having worked for twenty years in the field, I know that it progresses far too fast to allow an encyclopedic presentation, even when restricted to cooperation in human societies, which is a tiny fraction of all the cooperation encountered in biology.

Rather than trying to address all aspects, this book concentrates on one issue only, the reciprocity between self-interested individuals, and explores it for a small number of elementary types of interactions. The method is based on an evolutionary approach: more successful strategies become more frequent in the population. We neglect family ties, or neighborhood relations, or individual differences, or cultural aspects. It is best to state this self-limitation right at the beginning. I hope not to convey the impression that family ties, neighborhood relations, or individual aspects, etc., play no role in the evolution of cooperation and that it all reduces to self-interest; just as theoretical physicists writing a treatise on gravity do not imply, thereby, that other forces in the universe can be ignored. This being said, the current trend in economic life seems to lead away from nepotism, parochialism, and face-to-face encounters, and toward interactions between strangers in a well-mixed world.

The introduction (an entire chapter without any formulas) describes some of the most basic social dilemmas. Thinkers throughout the ages have been fascinated by the topic of self-regarding vs. other-regarding behavior, but the use of formal models and experimental games is relatively recent. Ever since Robert Trivers introduced an evolutionary approach to reciprocity, the Prisoner's Dilemma game serves as a kind of model organism to help explore the issue. But other games, such as the

Ultimatum, are quickly catching up. The most gratifying aspect of this development is the close connection between theoretical and experimental progress.

The second chapter provides a self-contained introduction to evolutionary game theory, stressing deterministic dynamics and stochastic processes, but tying this up with central notions of classical game theory, such as Nash equilibria or risk-dominance.

The third chapter provides a detailed discussion of repeated interactions, such as the Prisoner's Dilemma or the Snowdrift game, which allow exploration of direct reciprocity between the same two players meeting again and again. In particular, simple strategies based on the outcome of the previous round (such as *Tit for Tat*) or implemented by finite automata (such as *Contrite Tit for Tat*) offer a wide range of behavior.

The fourth chapter is devoted to indirect reciprocity. Here, players interact at most once, but they are aware of the past behavior of their one-shot partner. This introduces topics such as moral judgment or concern for reputation. Strategies based on the assessment of interactions between third parties allow the emergence of types of cooperation immune to exploitation, because they are channeled towards cooperators only.

The fifth chapter deals with the Ultimatum and the Trust game. Such games allow one to tackle the issues of fairness and trust, and provide, as a kind of side benefit, a framework for analyzing the roles of positive and negative incentives. Again, reputation plays an essential role for cooperation to emerge.

The sixth chapter turns from interactions between two players to interactions within larger groups. In so-called Public Goods games, defection can be suppressed by rewards or sanctions. Such incentives, properly targeted, allow reciprocation in mixed groups of cooperators and defectors. An intriguing aspect concerns the role of voluntary, rather than compulsory, participation in the team effort. Coercion emerges more easily if participation is optional.

The short seventh chapter, finally, deals briefly with some of the many issues that were neglected, such as nepotism, localized interactions, or group selection.

Needless to say, this book owes much to my colleagues, many of whom read draft chapters and provided comments. In particular, I want to thank Christoph Hauert, Arne Traulsen, Hannelore De Silva (formerly Brandt), Hisashi Ohtsuki, Satoshi Uchida, Ulf Dieckmann, Tatsuya Sasaki, Simon Levin, Ross Cressman, Yoh Iwasa, Silvia De Monte, Christoph Pflügl, Christian Hilbe, Steve Frank, Simon Gächter, Benedikt Hermann, Dirk Semmann, Manfred Milinski, and Josef Hofbauer. Most of all, I am indebted to Martin Nowak, without whom this book could never have been written.

The Calculus of Selfishness

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	10
	if I refuse to contribute	40	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

this structure are also known as *Staghunt* games. A fictitious population will evolve towards a state where all play the same strategy. The outcome depends on the initial condition. It is easy to see that if there are more than ten percent *TFT* players around, they will succeed. If the probability of another round is close to 1, i.e., if the expected number of future rounds is large, then even a small percentage of reciprocators suffices to overcome the defectors.

The computer simulations show, however, that a *TFT* regime is not the "end of history." Indeed, *AllC* players can invade, since in a *TFT* world, they do as well as the retaliators. If a small minority of *AllC* players is introduced into a population where all residents play *TFT*, they will do just as well as the resident majority. In fact, under plausible conditions they even offer an advantage. Indeed, an unconditional strategy seems easier to implement than a conditional strategy. More importantly, if a mistake occurs in an interaction between two *TFT* players, either because a move is mis-implemented or because it is misunderstood by the co-player, then the *TFT* players are caught in a costly sequence of alternating defections, in the relentless logic of "an eye for an eye." In computer simulations, such mistakes can be excluded, but in real-life interactions, they cannot. Mis-implementing a move or misunderstanding the co-player's action is common. An *AllC* player is much less vulnerable to errors: a mistake against a *TFT* player, or against another *AllC* player, is overcome in the very next round.

If individual-based simulations are life-like enough to allow for occasional errors, then a *TFT* regime is unlikely to last for long; less stern strategies such as *AllC* can spread. But once a substantial amount of *AllC* players is around, then *AllD* players can take over. The evolutionary chronicles of artificial populations involved in repeated interactions of the Prisoner's Dilemma type are fascinating to watch. The outcome depends in often surprising ways on the range of strategies tested during the long bouts of trial and error provided by the individual-based simulations. One frequent winner is *Pavlov*, a strategy that begins with a cooperative move and then cooperates if and only if, in the previous move, the co-player choose the same move as oneself. In chapter 3, we shall analyze some of the game theoretical aspects behind individual-based simulations.

1.8 THE CHAMPIONS OF RECIPROCITY

The computer tournaments led to a wave of research on reciprocity. But how much of it relates to the real world, as opposed to thought experiments? If *Tit for Tat* is so good, it should be widespread among fish and fowl. Evolutionary biologists and students of animal behavior uncovered a handful of candidates, but no example was universally accepted. It is difficult, in the wild, to make sure that the payoff values (which, in the animal kingdom, are expressed in the currency of reproductive success) do really obey the ordering given by equation (1.1). It is even more difficult to infer, from observing the outcome of a few rounds, which strategy was actually used. *TFT* is but one of countless possibilities. Moreover, many real-life collaborations offer plenty of scope for other explanations, for instance via kin-selection.

Today, after a few decades of this research, the net result is sobering. Beyond the realm of primates, there are few undisputed examples of *Tit for Tat*–like behavior. On the other hand, an overwhelming body of evidence proclaims that humans are, far and wide, the champions of reciprocity. This is not only clear from a huge amount of psychological tests and economic experiments. Brain imaging seems to support the view that part of our cortex is specialized to deal with the ceaseless computations required to keep count of what we give and what we receive, and to respond emotionally to perceived imbalance. Moreover, humans have an extraordinary talent for empathy—the ability to put oneself into another's shoes. Not only do we have an instinctive propensity to imitate another person's acts, we also are able to understand the intentions behind them.

For human nature, retaliation comes easy. The impulse is so strong that little children kick back at inanimate objects that hurt them. More importantly, we empathize with strangers interacting with each other, even as mere bystanders, as so-called *third parties*. This opens up the field of indirect reciprocation.

1.9 ENTER THE THIRD PARTY

You may know the old story about the aged professor who conscientiously attends the funerals of his colleagues, reasoning that "if I don't come to theirs, they won't come to mine." Clearly, the instinct of reciprocation is misfiring here. On second thought, it seems likely that the funeral of the professor, when it comes, will indeed be well-attended. His acts of paying respect will be returned, not by the recipients, but by third parties. This is indirect reciprocity.

In direct reciprocity, an act of helping is returned by the recipient. "I'll scratch your back because you scratched mine." But in indirect reciprocity, an act of helping is returned, not by the recipient, but by a third party. "I'll scratch your back because you scratched somebody else's." This seems much harder to understand. Nevertheless the principle suffices, so it seems, to keep cooperation going—or more precisely, to keep it from being exploited, and thereby ruined.

Indeed, an exploiter will gladly accept help without ever giving anything in return. If all do this, cooperation has vanished. Therefore, such exploitation should be repressed. The obvious way to do this is to withhold help from those who are known to withhold help. This channels cooperation towards the cooperators. But a moment's reflection shows that the principle is not consistent: if you restrain from helping an exploiter, you may be perceived by third parties as an exploiter yourself, and suffer accordingly. But we shall see in chapter 4 that indirect reciprocation can nonetheless hold its own.

If third parties can distinguish between a justified refusal to help an exploiter, and an unjustified refusal, then those who refuse to help exploiters run no risk of being seen as exploiters themselves. Bystanders must be able to assess actions as justified or not, i.e., as good or bad, even when they are not directed at themselves.

A closer investigation reveals that there are many possible assessment norms. Some work better than others. All require a considerable amount of information about the other members of the population. The faculty to process such information

may have evolved in the context of direct reciprocity already. It is certainly helpful, before you launch into a series of iterated games, to know how your prospective partners behaved towards their previous co-players. In this sense, indirect reciprocity "may have emerged from direct reciprocity in the presence of interested partners," in the words of evolutionary biologist Richard Alexander. But whereas direct reciprocity requires repetition, indirect reciprocity requires reputation. In the former case, you must be able to recognize your co-players; in the latter, you must know about them. "For direct reciprocity, you need a face; for indirect reciprocity, you need a name" (David Haig).

Subscribers to eBay auctions are asked to state, after each transaction, whether they were satisfied with their partner or not. The ratings of eBay members, accumulated over twelve months, are public knowledge. This very crude form of assessment seems to suffice for the purpose of reputation-building, and seems to be reasonable proof against manipulation. Other instances of public score-keeping abound in social history: a cut thumb signified a thief, a shaved head told of a fallen woman, a medal announced a hero. Reputation mechanisms have also played an important role in the emergence of long-distance trade.

If the community is small enough, direct experience and observation are likely to be sufficient to sustain indirect reciprocity. In larger communities, individuals often have to rely on third-party knowledge. Gossip must always have been the major tool for its dissemination. It may well be that our language instinct and our moral sense co-evolved.

1.10 MORAL SENTIMENTS AND MORAL HAZARDS

The role of moral judgments in everyday economic decisions was well understood by Adam Smith, who wrote his book on *The Theory of Moral Sentiments* even before turning to *The Wealth of Nations*. Later generations of economists tended to neglect the issue of moral emotions. But today, it is generally recognized that our "propensity to trade, barter, and truck" requires, first and foremost, trust. Trust has been hailed as a "lubricant of social life." Different communities operate on different levels of mutual trust. A firm moral basis for economic interactions and a consensual "rule of law" appear to be major indicators for the wealth of nations, more important than population size or mineral resources.

The human propensity to trust is well captured in the so-called Trust game. This is built upon the Donation game: in the first stage, the Donor (or Investor) receives a certain endowment by the experimenter, and can decide whether or not to send a part of that sum to the Recipient (or Trustee), knowing that the amount will be tripled upon arrival: each euro spent by the Investor yields three euros on the Trustee's account. In the second stage, the Trustee can return some of it to the Investor's account, on a one-to-one basis: it costs one euro to the Trustee to increase the Investor's account by one euro. This ends the game. Players know that they will not meet again. Clearly, a selfish Trustee ought to return nothing to the Investor. A selfish Investor ought therefore to send nothing to the Trustee. Nevertheless, in real experiments, transfers are frequent, and often lead to a beneficial outcome for both players. The

Trust game is analyzed in chapter 5, where it is shown that, unsurprisingly, concerns for reputation play a vital role.

Many real-life economic interactions contain elements of the Trust game. For instance, if I entrust money to a fund manager, I expect a positive return; and the fund manager also expects a benefit. The most important asset of a fund is its good reputation. A banker who fails to return the money will meet double trouble. On the one hand, the persons who entrusted him with their money will insist on getting it back; on the other hand, no new clients will be willing to trust him with their earnings. Both direct and indirect reciprocity are at work.

Economists and social scientists are increasingly interested in indirect reciprocity because one-shot interactions between far-off partners become more and more frequent in today's global market. They tend to replace the traditional long-lasting associations and long-term interactions between relatives, neighbors, or members of the same village. A substantial part of our life is spent in the company of strangers, and many transactions are no longer face-to-face. The growth of e-auctions and other forms of e-commerce is based, to a considerable degree, on reputation and trust. The possibility to exploit such trust raises what economists call moral hazards. How effective is reputation, especially if information is only partial?

Evolutionary biologists, on the other hand, are interested in the emergence of human communities. A considerable part of human cooperation is based on moralistic emotions, such as, for instance, anger directed towards cheaters, or the proverbial "warm inner glow" felt after performing an altruistic action. It is intriguing that humans not only feel strongly about interactions that involve them directly, but also about actions between third parties. They do so according to moral norms. These norms are obviously to a large extent culture-specific; but the *capacity* for moral norms appears to be a human universal for which there is little evidence in other species.

It is easy to conceive that an organism experiences as "good" or "bad" anything that affects its own reproductive fitness in a positive or negative sense. Our pleasure in eating calorie-rich food or experiencing sex has evolved because it heightens our chances of survival and reproduction. In the converse direction, disgust, hunger, and pain serve as alarm signals helping us to avoid life-threatening situations. The step from there to assessing actions between third parties as "good" or "bad" is not at all obvious. The same terms "good" and "bad" that are applied to pleasure and discomfort are also used in judging interactions between third parties: this linguistic quirk reveals an astonishing degree of empathy, and reflects highly developed faculties for cognition and abstraction.

1.11 ULTIMATUM EXPERIMENTS

A series of economic experiments documents that indirect reciprocity works. The more the players know about each other, the more they are likely to provide help to each other. There seems clear evidence for the player's concern with their own reputation. But interestingly, many players also tend to help, although to a lesser degree, when they know that nobody can watch them and that their action will not

affect their reputation. Moreover, they are more likely to give help if they have previously received help. This is difficult to explain through self-interest. It could be the outcome of a maladaptation. If somebody holds a door open for you, then you are more likely to hold the door open for the next person, motivated by a vague feeling of gratitude. It may well be that similar reflexes of misdirected reciprocity operate in other social and economic contexts.

A particularly revealing light on our propensity to empathize with others is provided by the Ultimatum game. In this experiment, two anonymous players are randomly alloted the role of Proposer and Responder. The Proposer is then given 10 euros, and asked to divide that amount between the two players, subject to the Responder's acceptance. Thus if the Responder accepts the proposed split, then the money will be shared accordingly, and the game is over. But if the Responder rejects the offer, then the game is also over; the experimenter withdraws the 10 euros, and both players receive nothing. This is it: no haggling, and no second round.

It seems obvious that the Responder should accept any positive offer, since this is better than nothing. Accordingly, a selfish Proposer should offer only a minimal share. In real experiments, however, most players offer a fair split—something between forty and fifty percent of the total. On the few occasions that less than twenty percent is offered, the Responder usually refuses. Proposers seem to anticipate this.

In most cases, refusals are correlated with angry feelings. Brain imaging shows that unfair offers elicit activity in two brain areas: one is in the left frontal part of the brain, which is usually associated with rational decisions, while the other is much deeper, in the striatum, which is linked with emotional responses. The tug of war between these two parts of the brain corresponds to the tension between (a) accepting the low offer, on the grounds that it is better than nothing, and (b) telling the unfair Proposer to go to hell. The intensity of the brain activities in the two locations foretells the decision, even before the Responder is aware of it.

The Ultimatum game experiment has been repeated many times. A large number of variants have been explored. For instance, if the Proposer is a computer, the Responder feels no qualms in accepting a small offer. If a game of skill (rather than the toss of a coin) decides who of the two players is going to be the Proposer, then smaller offers are more likely to be accepted: it is as if the Proposer had earned the right to keep a larger part of the sum. Furthermore, if several Responders compete, the Proposer knows that a small offer has a good chance of being accepted.

1.12 FAIRNESS NORMS

An extensive research program has used the Ultimatum game to study fairness norms in many small scale societies, including hunter-gatherers, nomads, slash-and-burn farmers, etc. The average offer varies between cultures. Remarkably, offers in large cities are among the fairest; Mother Nature's son is not always as noble as a city slicker or even an economics undergraduate. But the average offer is always far from the theoretical minimum. Norms of fairness seem wide-spread, maybe universal. How did they emerge?

Again, one possible explanation relies on reputation. Once it becomes known that you reject unfair offers, people will think twice before proposing them to you. The long term benefit of rejecting the offer may well outweigh the loss, which is all the smaller, the smaller the share you have been offered. In chapter 5, a simple mathematical model reveals how concerns for reputation can lead to the establishment of fairness norms. Paradoxically, this works only if Proposers who, ordinarily, are willing to offer a fair share, do occasionally yield to the temptation of offering less if they can get away with it. It is thus precisely when fairness norms are not hard-wired, and may be overcome by the opportunistic urgings of selfishness, that these norms are upheld in the population.

What have real experiments (as opposed to individual-based computer simulations) to say about this? It is easy to set up two distinct treatments of the Ultimatum game, each with a large population of anonymous test subjects who are randomly paired. In one treatment, players play the game for ten rounds (always against different co-players, of course) and nobody knows anything about the outcome of the previous rounds. In the other treatment, the outcomes are known to all. It is obviously only in the second treatment that players can hope to establish a reputation for rejecting small offers. The outcome is clear: the unfair offers tend to be considerably rarer. It is as if the Proposers anticipate that Responders fear to get exploited if it becomes known that they have meekly consented to a trifling share.

If Responders, in the Ultimatum game, reject an unfair offer, they have every interest in letting this be known to others. Under natural circumstances, an emotional response is likely to attract some attention. Anger is loud.

This being said, the fact remains that Ultimatum offers are often fair even if players know that the outcome will be kept secret. This seems puzzling. But it could well be that the players' subconscious is hard to convince that nobody will ever know. In our evolutionary past, it must have been exceedingly difficult to keep secrets from the small, lifelong community of tribal members and village dwellers in which our ancestors lived. Moreover, the belief of an overwhelming majority in a personal god watching them day and night shows that the feeling of being observed is deep-rooted and wide-spread.

Psychologists have devised ingenious experiments to document that our concern of being observed is easily aroused. For instance, players sitting in a cubicle in front of a computer are strongly affected by the mere image of an eye on the computer screen. They know that the eye is purely symbolic, but nevertheless they react to it. In another wonderfully simple experiment, the mere picture of eyes on a cafeteria wall next to the "honesty box" in a British university department sufficed to raise the amount staff members paid for coffee and cookies by more than two hundred percent. Obviously, it is easy to trigger a concern about being watched. And it is worth emphasizing that in our species, the eyes are uniquely revealing: due to the white color around the iris, the direction of their gaze is particularly noticeable. Incidentally it seems that test persons react the same, whether one or several persons are watching. This shows that they believe, at least subconsciously, that news will spread through gossip. One witness is enough.

1.13 PUBLIC GOODS GAMES

The games considered so far, such as Prisoner's Dilemma, Snowdrift, Trust, or Ultimatum, are two-person games. But many economic interactions involve larger groups of actors. The notion of reciprocation becomes problematic, in such cases. If your group includes both cooperators and defectors, whom do you reciprocate with? This introduces a new twist to social dilemmas.

So-called Public Goods games offer experimental instances of such dilemmas. Here is a typical specimen of such an experiment: Six anonymous players are given 10 dollars each, and are offered the opportunity to invest some of it in a common pool. The players know that the content of the common pool will subsequently be tripled by the experimenter, and that this "public good" will then be divided equally among all six players—irrespective of the amount that they contributed.

Obviously, all players are well off if they fully invest their 10 dollars. They receive 30 dollars each. But if one player invests nothing, and the others contribute fully, then each of the six players receives 25 dollars back from the public good; the defector, who contributed nothing, and thus kept the initial 10 dollars, ends up with a net sum of 35 dollars, 10 dollars more than the others.

For each dollar invested, only 50 cents return to the contributor. A selfish income-maximizer ought to invest nothing. But if all players do this, they have missed a first-class opportunity to increase their stocks.

In real experiments, most players invest on average half their initial amount, or even more. There are considerable variations among the individual contributions, but many players seem to hedge their bets. However, if the game is repeated for a few rounds, the contributions decline from round to round, and may end up at zero. The mechanism seems clear. If players notice that they have contributed more than others, they feel exploited, and reduce their future investments. But this causes other cooperators to feel exploited, and they reduce their contribution in turn. Cooperation goes down the drain.

In the repeated Prisoner's Dilemma game, a strategy like *Tit for Tat* allows one to retaliate against defectors. Such a reciprocating strategy loses its clout in a repeated Public Goods game. Indeed, by withholding your contribution, you hit friend and foe alike: your response is not directed against defectors only, but affects all the participants of the Public Goods game.

In economic life, similar interactions based on joint efforts, or joint investments, abound. This social dilemma is often described as multi-person Prisoner's Dilemma, or Free-Rider problem, or Tragedy of the Commons. A commons is a piece of grazing land that can be used by all villagers. The tragedy of the commons is due to the fact that it is usually over-exploited, and therefore ruined through overgrazing. Today, there are not many commons left, but the tragedy is still with us: the oceans are our new commons. On a smaller scale, the tragedy can be seen in most communal kitchens. Joint enterprises and common resources offer alluring prospects for cheaters and defectors.