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 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

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

		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

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.

1.14 PUNISH OR PERISH?

If you try riding free on public transportation, or dodging taxes, or littering parks, you run the risk of being caught and fined. Many judicial and legal institutions, as well as moral pressure, aim at keeping our contributions up. Thus the free-rider problem has an obvious solution: cooperation can be bolstered through incentives, by punishing or rewarding individual players, conditional on their behavior. However, legal institutions require a fairly advanced society.

It turns out that in the absence of such institutions, individuals are often willing to make the job of sanctioning their own.This has been neatly demonstrated by a series of experiments. After a round of the Public Goods game, players are told what their co-players contributed, and are given the opportunity of punishing them. If players are punished, this means that a fine of three dollars is deducted from their account. This fine is collected by the experimenter, and does not end up in the pockets of the punishing player. On the contrary, the punishing player must pay one dollar to inflict the punishment: this can be viewed as a fee which has to be paid to the experimenter. The fee is meant to reflect the fact that punishing another individual is usually a costly enterprise: it demands time and energy, and often enough entails some risk.

In the economic experiments, players are often willing to punish, despite the cost to themselves. This seems to be anticipated by most participants. The average level of contributions is higher, with the threat of punishment, than without. Most significantly, if the game is repeated for several rounds (each consisting of a Public Goods game followed by an opportunity for meting out punishment), then the contributions increase from round to round, up to remarkably high levels, see figure 1.1. Punishment obviously boosts the level of cooperation.

Why do people engage in costly punishment? The first explanation is obvious. By punishing defectors, one can hope to reform them. Thus punishers can expect to recoup their fee in the following rounds, through the heightened contributions of the castigated players. But this appears to be only part of the answer. Indeed, in a variant of the game requiring a large population of test persons, the Public Goods groups of six players are newly formed between rounds, and players know that they will never meet a co-player twice. By inflicting punishment, they can possibly turn a defector into a cooperator. However, punishers know that the future contributions of such an improved player will exclusively benefit others. Punishment appears as an altruistic act.

This is a stunning outcome. Without sanctions, the public good, i.e., the tripling of the endowment, is not realized. With sanctions, it is, although selfish reckoning prescribes that costly punishment should not be delivered. In the absence of institutions, some players are willing “to take the law into their own hands,” which is also known as “peer-punishing.”

What motivates players to punish defectors? Probably, we need not invoke any drive beyond the prevailing tendency to reciprocate. If the players themselves feel exploited, direct reciprocation is at work. If others are exploited, it is indirect reciprocation: humans are often willing to retaliate on behalf of third parties.

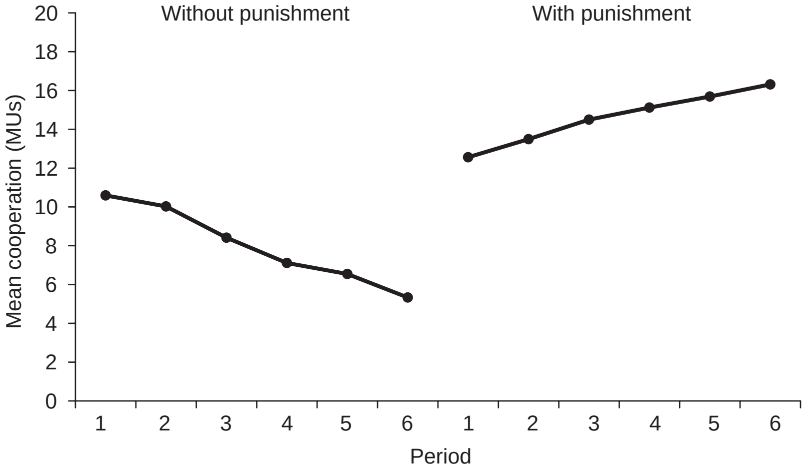

Figure 1.1 Public Goods and punishment. In each of the twelve rounds of the Public Goods game, groups of four players are formed (out of a population of 240 players). The players receive 20 monetary units (MU) per round, and have to decide how much of it to invest, knowing that their contributions will be multiplied by two and divided equally among the four participants. In the treatment "with punishment," players can fine their co-players; fines are collected by the experimenter. Imposing a fine of three monetary units costs a fee of one monetary unit. Players know that they encounter a co-player only once. Shown is the average contribution to the public good in each round. (After Fehr and Gächter, 2002.)

1.15 SECOND-ORDER FREE-RIDING

It is clear that in a population consisting of players ready to punish exploiters, defection makes no sense. The gain from not contributing is more than off-set by the expected fines. Defectors would have to bear the full brunt of punishment from the majority. If punishers (i.e., players who contribute, and impose fines on those who do not contribute) are established in a population, they can resist defectors and uphold cooperation.

But a population of punishers can be subverted by players who contribute, but do not punish. Newcomers of that type do just as well as the resident punishers and thus can slowly spread through random fluctuations. In fact, if occasionally some defectors enter the population, to be promptly assailed by the punishers, then the newcomers would do better than the punishers, by economizing on the cost of punishment. This new type is a second-order exploiter, free-riding on the sanctions delivered by the punishers. Hence, this type will spread: and this means that eventually, there will be too few punishers around to keep the defectors at bay.

Sanctioning can be seen as a service to the community, i.e., a public good. In the long run, second-order exploiters sabotage the enforcement of contributions to the Public Goods game, and therefore both types of contributors—the punishers and the second-order exploiters themselves—will eventually be displaced by defectors.

A remedy coming to mind is “second-order punishment.” It consists in punishing not only the “first-order exploiters” who fail to contribute, but also the “second-order exploiters” who contribute, but fail to punish. However, this could in turn give rise to “third-order exploiters” and so on. If punishers of a sufficiently high order dominate the population, there will be few defectors, and therefore few occasions for the lower-order punishers to reveal their limitations to their sterner brethren. Hence, their number can increase through random fluctuations, thus eroding the system.

It seems that reputation, once more, can come to the rescue. Players are less likely to yield to the temptation to cheat if they know that their group includes some punishers. Thus, there is an advantage in being known to react emotionally against exploiters. This will be analyzed in chapter 5. The situation is quite similar to the Ultimatum game. In fact, a Responder who refuses an unfair offer is effectively punishing the Proposer. The more unfair the offer, the less is the cost to the punisher, and the heavier the fine to the punished player.

A similar mechanism operates with positive incentives. If players, after a Public Goods round, are given the possibility to reward the high contributors, at a cost to themselves, they are able to promote the tendency to cooperation. Again, this system is threatened by those players who contribute, and thus benefit from rewards, but do not reward others. Such players are obviously free-riding at the expense of the rewarders, and can subvert the incentive-based system. But if rewarding players can acquire a reputation, they are more likely to experience high levels of cooperation in their group.

The similarity between reward and punishment stops at a point. In a society where everyone contributes, punishers have nothing to do, but rewarders have to dig deep into their pockets. In this sense, punishment is more efficient. Ideally, the mere threat suffices. In real experiments, however, one often finds that while fines do certainly increase cooperation, they may be so expensive that the average payoff in the group is actually lower than in the less cooperative groups playing the Public Goods game without punishment, at least during the first few rounds. Moreover, in many societies asocial punishment (i.e., the punishment of do-gooders) is frequent, and thus throws a spanner in the work of sanctions to uphold cooperation.

1.16 VOLUNTARY PARTICIPATION

Even granted that reputation can stabilize a population of punishers, there remains the problem of explaining how sanctioning can emerge. In a world of defectors, punishers would have to punish right and left. Their payoff would be low and their behavior unlikely to catch on.

This is different, however, if players are not obliged to participate in the Public Goods game, and can opt out of it if they wish. This situation seems natural enough. In many cases, individuals can decide whether to play the game or not. In town, you need not use public transportation: walking is fine. In a hunter-gatherer tribe, you need not join the big-game hunt, or the raiding party, if you suspect that the other participants are laggards. Collecting mushrooms or fruits can provide an option that makes you independent from the others. You need no assistance.

Suppose thus that there exists an alternative to participation in the joint effort, an alternative whose outcome does not depend on what the others are doing. We may then see the Public Goods game as a newly arising opportunity. A mammoth has moved into your valley. Will it pay to join the hunt? Participating in the common effort means effectively to bet on cooperation. We shall assume that if all participants contribute, engaging in the Public Goods game is more profitable than the alternative of not participating; but that if the other participants do not contribute, the Public Goods game is a waste of time that ought to be avoided. Searching for mushrooms is more promising, in that case.

Let us first consider this "optional Public Goods game" without punishment. In that case, the three strategies (to contribute, to defect, or to abstain) are superseding each other in a cyclic fashion, as in the familiar Rock-Paper-Scissors game. If the population consists mostly of cooperators contributing to the joint effort, then it is best, from the selfish viewpoint, to exploit them. But if most players switch to defection, then the Public Goods game is unprofitable and it is better not to participate at all. Finally, if most players are not participating, then cooperation is the best option. This last statement may seem surprising. But if few players are willing to participate, then most teams will be small, and in this case cooperators can do better, on average, than defectors, despite the fact that within each team, defectors do better than cooperators.

These Rock-Paper-Scissors cycles, from contributing to defecting to abstaining to contributing again, do not yet suffice to establish cooperation. In the long-term average, the payoff is not higher than the payoff for non-participants. But as we shall see in chapter 6, if the option of punishing the exploiters is added, then cooperation will be established for most of the time. This is a statistical result. Under stochastic fluctuations, punishers can be subverted after some time by second-order exploiters, and these in turn by defectors; but after such a break-down of cooperation, punishers re-emerge. In the long term, they dominate the population.

The outcome seems paradoxical. In interactions requiring a joint effort, cooperation based on coercion can emerge and prevail, but only if the participation is voluntary. If participation is compulsory, coercion fails and defectors win.

Several economic experiments support the validity of this theoretical conclusion. In Prisoner's Dilemma games and Public Goods games, cooperation is more likely to be achieved if players have the option to abstain from the game. In one particularly telling experiment, players from a large pool had the possibility to choose, between rounds, not whether to participate in a Public Goods game or not, but whether to play their Public Goods game with or without the punishment option. In the first round, most players decided against the version with punishment. This seems understandable. Nobody wants to be punished, and many people dislike punishing; but by the fifth or sixth round, almost all players had switched, on their own free will, to the version with punishment, and cooperated assiduously. They effectively "voted with their feet" for the threat of sanctions, understanding that it made cooperation more likely. This experiment looks almost like a morality play, illustrating the philosophy of the social contract.

1.17 THE GENTLE ART OF ABSTRACTION

How relevant are economic experiments? Often, their most striking aspect is a stark artificiality. They are remote from everyday experience.

This in itself need not be a weakness. Classical experiments in physics or physiology are equally remote from everyday life. Their aim is to probe nature, not to mimic it. It can be argued, for instance, that a major asset of the Ultimatum game consists in creating a situation that players have never encountered before. We have been exposed to haggling, to the rituals of offer and counteroffer, and to market competition. The barren "take it or leave it" alternative of the Ultimatum is profoundly alien to most players. By catching us on the wrong foot, the experimenter forces us to decide spontaneously, rather than rely on force of habit.

The anonymity under which most economic experiments are performed excludes all possible effects of relatedness, reputation, future interactions, or advertising. Anonymity is not a condition that humans have often encountered in their evolutionary past. Most of human evolution took place in small tribes and villages, with everyone knowing everything about everyone else. We have certainly not been adapted, in our evolutionary past, to transferring small sums of money under contrived rules to faceless strangers. It makes no sense to assume that Ultimatum games or Trust games, in their clinical sterility, have shaped our evolution. But human behavior is based on evolved traits, and by varying the treatments in economic experiments, we may hope to unveil these traits.

For instance, players who are allowed to briefly talk with each other, before engaging in a Prisoner's Dilemma game, are more cooperative. Moreover, they can predict very accurately, after a short conversation, whether their co-players will cooperate or not. Even without knowing which type of experiment is in store for them, they quickly pick up the relevant clues for summing up their partner. By varying the nature of their conversation, which can be face-to-face, via monitor, through a phone, or merely a brief visual contact, experimenters can hope to understand how we go about assessing strangers.

To give another example, players tend, as we have seen, to reject unfair offers more readily if they know that this becomes known to their future co-players. Nevertheless, even players who are assured that nobody will know about their decision frequently turn small offers away. It would be naive to overlook the possibility that even if players are convinced that nobody is watching, and have grasped the niceties of double-blind experiments, their subconscious may yet harbor some misgivings. We are far from completely understanding when and why subliminal factors can affect decision making. Players can strongly react to an appropriate cue even when knowing that reality does not back it up. An often mentioned example is the sexual arousal produced by erotic magazines.

Experimental game theorists know this, of course. They do not try to reproduce real life interactions, with their plethora of psychological and cultural effects, but aim to dissect the strategic situation down to the bones. Most economic interactions take place with innumerable side-conditions, among people bound by a plethora of ties of personal history and cultural constraints. Experiments must abstract from all these factors.

1.18 HUMAN BEHAVIOR RESUMED IN TWO SECONDS

In a similar spirit of self-imposed limitation, the mathematical models filling most of this book omit all psychological factors but one: selfishness. This by itself need be no severe restriction: most psychologists would agree that it is a good first approximation. (To quote Jonathan Haidt: “If you are asked to explain human behavior in two seconds or less, you might want to say ‘self-interest’.”). Some very interesting and plausible theoretical approaches assume that individual utilities include the utilities of other players—that equity, for instance, is deemed desirable—but this eminently psychological issue cannot be tackled here. It seems likely that our preferences emerged through evolution, and that a direct path led from the “selfish gene” to human kindness, but such a topic is way beyond the scope of this book, which merely explores, by mathematical means, how selfishness can overcome social dilemmas.

This is not meant to endorse the idea that our social interactions are governed by some “homo economicus” residing in our breast, who calculates strategies to maximize our own gain with cold rationality. According to John Maynard Keynes, economic decisions are often governed by “animal spirits and spontaneous optimism” and depend on “nerves and hysteria and even digestion and reactions to the weather.”

Emotions and instincts act as a system of heuristics to guide us through computations which vastly overtax our rational faculties. Similarly, tennis players manage to compute the trajectory of a ball with a speed and precision that no robot can match. The players work it out subconsciously, and it is doubtful whether any Wimbledon winner would become a better player by a course in physics. In an analogous way, we need no pen and paper to figure out our self-interest in practical life. Game theory may, like a course in physics, provide understanding, but it need not furnish recipes for success.

Just as we concentrate, in the following pages, on self-interest as guiding motivation, we will also purposely ignore the effect of social structure. Neglecting networks may be an even more serious distortion of real life than neglecting altruism. The short last chapter 7 provides a brief glimpse at some factors that are left out in all the preceding chapters: namely family ties, neighborhood effects, and group benefits.

The major part of this book thus deals with simple games of cooperation played by selfish individuals in well-mixed, and usually large, populations. This is an admittedly artificial scenario, but our world seems to evolve towards it. Is it the way of the future? It certainly was no part of our evolutionary past. Nothing prepared us for big city life, but we do have an uncanny talent for mixing with strangers and enjoying “the tumultuous sea of human heads,” like the nameless hero of Edgar Allan Poe’s short story, “The Man of the Crowd.”

1.19 FOOD AND MORALS

Most economic experiments use real money, and some critics say this is about all that is real about them. But in fact, a large amount of everyday economic cooperation involves no money at all. We can lend a hand, or provide some information, or share a meal: in each case the psychological feeling is different. To use money, in experiments, is a simple, clear-cut way to reduce framing effects that complicate the strategic issue.

Nevertheless, it is obvious that this way of standardizing outcomes can sometimes be seriously misleading. For instance, when you had been thinking through the alternatives of the Donation game in section 1.3, you probably felt uneasy about one scenario. If the other player trusts you, would you be willing to defect? Most people balk at that point. It usually feels bad to let another person down. The discomfort seems hardly worth the few extra dollars. Indeed, many actual experiments indicate that a majority of players are willing to cooperate. A comforting amount of people are "good-natured." But where does this good nature come from?

Similar questions are raised by the Ultimatum game. Most Proposers offer close to half of the sum and claim that it just seems the fair thing to do. Conceivably, they are fooling themselves, and are simply afraid that a lower offer may be rejected. But why do Responders reject a small offer? Most claim that they are angered by the obvious injustice of the unfair offer. Again, they possibly mistake their own motivation, and are simply anxious to avoid the reputation of being spineless wimps, a reputation that would harm them in the long run. These selfish imputations seem to fail, however, in a variant of the Ultimatum game, which is known as the Dictator game. In this variant the Proposer makes the offer, and the Responder has no say at all: "Dictators" can do as they like.

In the Dictator game, the offers are usually lower than with the Ultimatum game. Nevertheless, a substantial part of the Proposers offers a positive amount. It seems difficult, in this case, to dismiss "good nature." Proposers simply feel that to be generous makes them happy. If, in another twist, Responders in the Dictator game cannot reject the offer, but can write a short note to let the Responder know what they think of it, then offers jump to almost the same level as with the Ultimatum game. Obviously, people do not like to incur the wrath of others, even if that wrath is guaranteed to be completely ineffective. Meting out the purely symbolic punishment of a censorious message is not very costly in that case.

Are we simply afraid of being cursed? It has been argued that a strong motive for cooperation and moral behavior is the fear of punishment by supernatural spirits. Superstitious maladaptations are widespread, possibly because they strongly promote conformism and obedience—traits which often have some survival value.

If we enjoy sex and food, it is because such emotions promote our survival and reproduction. Similarly, our survival and reproduction depends on being successful cooperators, and this is why we enjoy being virtuous, and why we feel that revenge tastes sweet. Moralistic emotions—the warm inner glow of feeling kind, the anger directed at unfair persons, the guilt and shame after committing a reprehensible deed—are deeply anchored in our nature. Moral rules differ among cultures, but juveniles' ability to pick up prevailing norms and make them their own seems to be as much a part of universal human nature as juveniles' ability to pick up and speak the language of their community.

Not all morality is meant to promote altruism and cooperation. Norms of personal cleanliness and purity have a similar ethical status, without having an economic background. But a large part of moral norms serve ultimately to promote the ceaseless give and take that is such an essential part of human behavior. The German playwright Bertold Brecht wrote in his *Threepenny Opera*: “Food comes first, then morals.” It is exactly the reverse. Without morals, we could not subsist.

But fortunately, we need not end with homilies. For what they are worth, the simple models analyzed here also contain some more surprising lessons: for instance, that the instinct of revenge, frowned upon as base, can play a useful economic role by deterring defectors; or that our selfish urge to exploit others whenever we can get away with it, keeps retaliators in the population, thus boosting common welfare; or that the option to abstain from a team effort when it appears unpromising actually helps in enforcing team-wise cooperation. Again and again, we find that traits rendering individuals less than perfect uphold social cohesion. So even if you cannot always satisfy your selfish interests, you may find consolation in the thought that they are furthering the common good.

That human and all-too-human foibles and errors sustain cooperation is not new, by the way. It is known as Mandeville’s paradox. The author of the *Fable of the Bees* subtitled his work with the slogan: “Private Vices, Publick Benefits.” Private selfishness can promote the public good. The “invisible hand” performs surprising tricks.

1.20 REFERENCES

Basic texts on the evolutionary biology of cooperation can be found in the works of Hamilton (1996) and Trivers (2002), see also Trivers (1985), Frank (1998), and Nowak (2006a). Popular expositions are given by Dawkins (1989), Sigmund (1995), and Ridley (1997). For game theoretical descriptions of social dilemmas with minimal technical fuss, see Colman (1995), Binmore (1994), Sugden (1986), Ostrom (1990), and Skyrms (2004). Good surveys on social dilemmas can also be found in Dawes (1980), Cross and Guyer (1980), Heckathorn (1996), Kollock (1998), and Levin (1999). The Tragedy of the Commons and the dilemmas surrounding collective action were presented by Hardin (1968) and Olson (1965). A popular account of the Prisoner’s Dilemma is provided by Poundstone (1992). The Prisoner’s Dilemma first mention in a textbook goes back to Luce and Raiffa (1957), see also Schelling (1978).The first full book devoted to the game is by Rapoport and Chammah (1965). Rapoport submitted the Tit for Tat strategy to Axelrod’s tournaments (Axelrod 1984). Indirect reciprocity can be traced back to Alexander (1987) and Ellison (1994). It was modelled by Nowak and Sigmund (1998a,b), and early experimental tests are described in Wedekind and Milinski (2000) and Wedekind and Braithwaite (2002). The Snowdrift game is due to Sugden (1986), see also Doebeli, Hauert, and Killingback (2004). The Trust game was first proposed by Berg, Dickhaut, and McCabe (1995), the Ultimatum game by Güth, Schmittberger, and Schwarze (1982).The Ultimatum’s use for investigating small-scale societies is covered, e.g., in Henrich (2006). The role of reputation in economics is studied, e.g., in Kreps and Wilson (1982) or Kurzban, DeScioli, and O’Brien (2007). Observer effects were observed in Haley and Fessler (2005), Bateson, Nettle, and Roberts (2006) and Burnham and Hare (2007). The role of sanctions in Public Goods games was studied by Yamagashi (1986) and Fehr and Gächter (2000, 2002), see also Boyd and Richerson (1992), O’Gorman, Wilson, and Miller (2005), Gardner and West (2004) and Sigmund (2007). The troubling aspects of antisocial punishment have been uncovered by Herrmann, Thöni, and Gächter (2008). An experiment by Gürerk, Irlenbusch, and Rockenbach (2006) shows that players of Public Good games opt for the possibility of sanctioning defectors, but only after some experience. Positive and negative incentives to promote cooperation have been compared in many investigations, see e.g., Baumeister et al. (2001), Dickinson (2001), Andreoni, Harbaugh, and Vesterlund (2003), Walker and Halloran (2004), or Sefton, Shupp, and Walker (2007). The role of voluntary participation is studied by Orbell and Dawes (1993), Hauert et al. (2002a, 2002b), and Fowler (2005a). Extensive monographs on experimental economics and behavioral games are by Kagel and Roth (1995) and by Camerer (2003), see also Camerer and Fehr (2006) and, for experiments under more natural conditions, Carpenter, Harrison, and List (2005). Theoretical, sociological and psychological studies on ethical norms and morals are by Yamagishi, Jin, and Kiyonari (1999), Bendor and Swistak (2001), Price, Cosmides, and Tooby (2002), Ostrom and Walker (2003), Cose (2004), Kurzban and Houser (2005), Hauser (2006), and Haidt (2007). Human universals are treated in Brown (1991). Richerson and Boyd (2005) offer a unified treatment of cultural and biological evolution. The role of gossip is highlighted by Dunbar (1996) and Sommerfeld et al. (2007). A seminal text on the economic importance of emotional committment is by Frank (1988). The neural basis of emotions related to economic games is studied in Rilling et al. (2002), Sanfey et al. (2003), Fehr (2004), and de Quervain et al. (2004). The importance of reciprocity is stressed by Charness and Haruvy (2002) and Sachs et al. (2004). Bowles and Gintis (2002), see also Gintis et al. (2003, 2005), present an influential approach termed "strong reciprocity": for a critical view, see Burnham and Johnson (2005). Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) show how to interpret experimental outcomes by modifying utilities, so as to incorporate concerns for equity and fairness. There exists a huge literature on economic and social interactions in non-human primates, see e.g., de Waal (1996), Brosnan and de Waal (2003), Stevens, Cushman, and Hauser (2005), Silk (2006), or Warneken and Tomasello (2006). Various forms of punishment in biological communities are covered in Clutton-Brock and Parker (1995), Kiers et al. (2003), or Wenseleers and Ratnieks (2006); for other ways of repressing competition, see Frank (2003).

GAME ZOO: A BRIEF LEXICON OF TWO-PERSON GAMES

Many experimental two-person games are related to social dilemma issues. Typically, the players are anonymous, and are endowed with a certain amount of money beforehand (e.g., a show-up fee). They are asked to make their decision after having understood the rules of the game and being assigned to the role of Proposer and Responder (or Donor and Recipient).

Donation: in some sense, an atom of social interaction. The Donor decides whether to pay one dollar to give a benefit of three dollars to the Recipient.

Prisoner's Dilemma: the mother of all cooperation games is played in many variations. In one particularly transparent set-up, both players engage in a Donation game with each other. When players decide simultaneously, this is similar to a two-player Public Goods game. If both cooperate by sending a gift to the other, both gain two dollars. But sending a gift costs one dollar, so that the best reply to whatever the co-player decides is to send no gift (i.e., to defect). If both players defect, however, they gain nothing.

Ultimatum: the experimenter assigns a certain sum, and the Proposer can offer a share of it to the Responder. If the Responder (who knows the sum) accepts, the sum is split accordingly between the two players, and the game is over. If the Responder declines, the experimenter withdraws the money. Again, the game is over: but this time, neither of the two players gets anything.

Dictator: same as Ultimatum, except that the Responder cannot reject the offer.

Trust: in a first stage, the Proposer (or Investor) can give a certain benefit to the Responder (or Trustee), as in the Donation game. In the second stage, the Responder can decide how much of it to return to the Proposer. This is similar to the sequential Prisoner's Dilemma game (when first one player acts as Donor and then the other).

Repeated Prisoner's Dilemma: the two players interact for several rounds of the Prisoner's Dilemma. Usually, they are not told beforehand when the interaction will be over, so as to avoid "last round effects" (defection motivated by the fact that the co-player cannot retaliate in a one-shot Prisoner's Dilemma game).

Indirect Reciprocity: in a large population of players, two players are sampled at random and play the Donation game or the (one-shot) Prisoner's Dilemma game. This is repeated again and again. The players know that they interact only once, so that retaliation is impossible.

Snowdrift: two players each receive an endowment, on provision that they pay a fee to the experimenter that is lower than the endowment. They must decide whether they are willing to pay the fee or not, knowing that if both are willing, each of them pays only half.