

THE RATIONAL SELF-INTEREST OF RECIPROCITY:

ROBERT AXELROD AND THE EVOLUTION OF COOPERATION



KEVIN MCFARLANE



Robert Axelrod's book, *The Evolution of Cooperation*, was first published in 1984 by Basic Books of New York. It was then made available in this country in 1990 through Penguin books. The 1990 edition contains a Foreword by the biologist, Richard Dawkins, author of *The Selfish Gene*. In fact, the book was indirectly inspired by Dawkins himself. Axelrod is an American political scientist who wrote to Dawkins in the late 1970s inviting him to compete in a computer tournament to play "Iterated Prisoner's Dilemma" (more of that later). Dawkins did not compete but instead wrote for Axelrod an introduction to the distinguished evolutionary biologist W. D. Hamilton who, unknown to Axelrod, resided at the same university! Hamilton and Axelrod then collaborated to produce a technical paper, "The Evolution of Cooperation in Biological Systems", which eventually formed the basis for the current book. (That paper is reproduced as Chapter 5.)

PRISONER'S DILEMMA

Axelrod begins by asking how, in a world of self-seeking egoists, cooperation can emerge if there is no central authority to police their actions. He then explains, through the use of game theory applied to the so-called Prisoner's Dilemma, how cooperation can indeed emerge under suitable conditions. The beauty of the book is that, though many of the conclusions depend on mathematical analysis for their *rigorous* justification, Axelrod explains things so lucidly that one does not need to know the mathematics to understand the arguments. Indeed, all the mathematics is relegated to footnotes and appendices.

For the purposes of his analysis Axelrod makes the assumption that people are narrowly self-interested, that is, that all actors are concerned only with their own welfare. Since this assumption is false it might seem that Axelrod's analysis is of somewhat limited practical use. However, there are several comments we can make here.

(1) The assumption of narrow self-interest is the "worst-case" scenario regarding the possibility of cooperation. So the solution of the problem of cooperation for this case actually *solves* the problem of cooperation for the general case. Because altruistic behaviour by A towards B, where B accepts A's altruism, is *necessarily* cooperative.

(2) If a person's self-interest is defined as also encompassing altruism towards relatives and friends the problem shifts from that of inducing cooperation between individuals to that of inducing cooperation between groups. The group, for example, the family or the state, can thus be regarded as the unit of self-interest *vis-à-vis* other groups. Then Axelrod's analysis applies at a higher level of abstraction.

(3) Altruism *within* a group does not imply altruism *at all times*. In other words, some of our actions will be narrowly self-interested and some will be altruistic even within groups of families or friends. (In reality, *most* of our actions *are* narrowly self-interested but among relatives and friends less of our actions are narrowly self-interested than among strangers or enemies.)

Sociological Notes No. 20

ISSN 0267-7113

ISBN 1 85637 264 2

An occasional publication of the Libertarian Alliance, 25 Chapter Chambers, Esterbrooke Street, London SW1P 4NN
www.libertarian.co.uk email: admin@libertarian.co.uk

© 1994: Libertarian Alliance; Kevin McFarlane.

Having previously worked as an engineer in the offshore oil industry, Kevin McFarlane is currently a Product Specialist for a scientific software company.

The views expressed in this publication are those of its author, and not necessarily those of the Libertarian Alliance, its Committee, Advisory Council or subscribers.

Director: Dr Chris R. Tame

Editorial Director: Brian Micklethwait

Webmaster: Dr Sean Gabb

FOR LIFE, LIBERTY AND PROPERTY



Axelrod begins his analysis by explaining the so-called *Prisoner's Dilemma* game, due originally to Merrill Flood and Melvin Dresher in about 1950. In this game there are two players who are awarded differing points according to whether they cooperate or defect. (To defect means to not cooperate.) The game works like this.

Let the two players be A and B. If A and B both cooperate they both get 3 points. This outcome for both players is called R, the reward for mutual cooperation.

If A cooperates but B defects then A gets 0 points and B gets 5 points. A's outcome is called S, the sucker's payoff. B's outcome is called T, the temptation to defect. So B gets a higher score if he defects while A cooperates than he does if they both cooperate. This is like A doing something for B but B doing nothing for A. The same results apply for A and B interchanged.

Finally, if A and B both defect they get 1 point. This outcome for both players is called P, the punishment for mutual defection.

HOW COOPERATION GETS STARTED

Why the dilemma? The dilemma arises when we consider what each player should do when he does not know what the other player is going to do.

For example, consider the game from player A's perspective.

First, assume that B will cooperate. Then:

- (1) if A cooperates A will get 3 points,
- (2) if A defects A will get 5 points.

Therefore, A concludes that it pays to defect if B cooperates.

Now assume that B will defect. Then:

- (1) if A cooperates A will get 0 points,
- (2) if A defects A will get 1 point.

Therefore, A concludes that it pays to defect if B defects.

So *regardless* of whether B cooperates or defects A is better off if he defects.

But, of course, since the situation is symmetrical, B must reach exactly the same conclusion as A. So the outcome of the interaction must be that *both* will defect. Both will get 1 point. Yet if only both had cooperated each would have got 3 points. Individual rationality leads to a worse outcome than is possible. Hence, the dilemma.

There is a technical reason for the specific magnitudes of the points assigned to the various outcomes, which I will not discuss here. But that the points scheme is unilateral defection > mutual cooperation > mutual defection > unilateral cooperation is quite a plausible assumption to make. (">" is the mathematical symbol for "greater than".)

The basic problem then is, given this dilemma, how can cooperation ever get going in a world of self-seeking egoists? The answer is that cooperation *can* get started *if* the players expect to meet repeatedly in the future. In other words, if we analyse the situation of the *iterated* Prisoner's Dilemma a very different picture can emerge.

Axelrod organised several computer tournaments in which the participants' computer programs were to play the game of iterated Prisoner's Dilemma on a round-robin basis, i.e., every program was to play every other program and was also to play against a copy of itself. The winner was to be the program which amassed the greatest number of points summed over all interactions. During a game each program was to have available to it the history of the interactions so far. This allowed for the possibility of more or less complex playing strategies. Axelrod ran a preliminary tournament and two slightly different rounds of a main tournament. The program which won the main tournaments was the simplest and shortest program of all. It was called TIT FOR TAT. This strategy also came second in the preliminary tournament and first in a variation of the preliminary tournament.

TIT FOR TAT follows the policy of cooperating on the first move and then doing whatever the other player did on the previous move. In other words, if the other player defects on the first move, or any other, TIT FOR TAT immediately retaliates by defecting. But whenever the other player cooperates TIT FOR TAT responds by cooperating.

ROBUSTNESS, NICENESS, COLLECTIVE STABILITY

Axelrod devotes much of his discussion to an analysis of these games of Prisoner's Dilemma, explaining why certain strategies are more successful than others and also examining their robustness. *Robustness* is the property of doing well in a wide range of possible environments, i.e., in environments where the mix of other strategies varies. It turns out that TIT FOR TAT scores top marks here as well. Interestingly, between the first and second tournaments Axelrod made available to everyone the results of the first tournament and also explained why TIT FOR TAT was so successful. But *still* no-one was able to improve on the strategy.

Axelrod then applies these lessons to practical situations as diverse as the behaviour of US senators, the Live-and-Let-Live policy of trench warfare in the First World War and the behaviour of biological systems.

The broad conclusions which Axelrod draws from his analysis are these.

- (1) Cooperation can get started even in a world of unconditional defection (everyone following a policy of always defecting, called ALL D). It can evolve from small clusters of individuals who base their cooperation on reciprocity and have even a small proportion

of their interactions with each other. But it cannot emerge if such individuals are too scattered and have a negligible proportion of their interactions with each other. In other words, there is a threshold fraction of interactions with like-minded strategies below which a benevolent or “nice” strategy such as TIT FOR TAT cannot do better than a “nasty” strategy such as ALL D. But that fraction can be as low as 5% for the particular parameters used by Axelrod. So if TIT FOR TAT has only 5% of its interactions with other TIT FOR TATS it can invade a population of ALL Ds. In other words, TIT FOR TAT’s average scores will be higher than those of ALL D.

(2) A strategy based on reciprocity can thrive in a world where many different kinds of strategies are being followed (robustness).

(3) Cooperation, once established, can protect itself from invasion by less cooperative strategies (also, robustness).

In order to quantify *robustness* Axelrod develops the concept of *collective stability*. Imagine a population of individuals, interacting between themselves, where all use the same strategy. Then suppose a newcomer comes along with a different strategy. The newcomer is said to invade the native population if the newcomer gets a higher score with a native than a native gets with another native. Since the newcomer is a sole individual this is equivalent to the newcomer getting a higher average score than the population average. A strategy is said to be *collectively stable* if no other strategy can invade it in this way.

This concept is useful in biological systems where, in particular ecosystems, variations in evolved strategies can be more or less successful (in terms of reproductive success) depending on what the predominant strategy is. (The biologist John Maynard Smith developed the similar but slightly different concept of the *evolutionarily stable strategy* in the 1970s.)

THE FUTURE

Whether a strategy is collectively stable or not is critically dependent on whether invading strategies arrive as isolated individuals or in clusters where they have a finite proportion of interactions among themselves.

Using TIT FOR TAT as the archetype of a cooperative strategy Axelrod derives two major results in connection with collective stability.

The first is related to point (3) above. A population of individuals interacting with each other using TIT FOR TAT cannot be invaded by any other strategy provided that the future is important enough. In other words, given that future interactions are *less important* than present interactions, the future interactions must be of *sufficiently great* importance in order for TIT FOR TAT to be a non-invadeable or collectively stable strategy.

The decreasing satisfaction to be had from increasingly distant future events is of course a well-known feature of economic theory, the phenomenon of so-called “time preference”. Axelrod provides a quantitative measure of the importance of the future in relation to the iterated Prisoner’s Dilemma but we need not concern ourselves with this here.

The second major result is related to point (1) above. TIT FOR TAT can invade a population of individuals using ALL D (all defect) provided that it has as little as 5% of its interactions with other TIT FOR TATS. (The reason for focusing on ALL D is that this is the least cooperative strategy, by definition, since it is following a rule which says: never cooperate.)

ALL D is in fact also collectively stable but only when other strategies try to invade it as lone individuals. This is because, in a population of All Ds, lone individuals have no-one with whom they can reciprocate cooperation. So All D is collectively stable against lone individuals but not against clusters of individuals. But TIT FOR TAT is collectively stable against *both* lone individuals *and* clusters of individuals.

SOME EXAMPLES

Axelrod then shows how these theoretical arguments can be applied to various social and biological settings. One example he describes is the Live-And-Let-Live system of trench warfare during World War 1. This is a particularly illuminating example since here we had cooperation between groups who were most definitely *not* supposed to cooperate since they were at war with each other! Axelrod explains how the conditions of trench warfare met the technical conditions of the Prisoner’s Dilemma, i.e., how the points for each pair of opposing battalions were arranged so that unilateral defection > mutual cooperation > mutual defection unilateral cooperation.

What made trench warfare different from most other combat was that the same small units faced each other in immobile sectors for extended periods of time. We have seen above that it is this fact that makes the TIT FOR TAT strategy viable. Axelrod describes how both sides did indeed follow a TIT FOR TAT strategy, which meant that both sides tended to cooperate but would respond to a defection by the other side. This behaviour tended to occur despite the best efforts of the opposing high commands to prevent it.

Axelrod then considers the evolution of cooperation in biological systems, where different members of the same, or different, species often cooperate. One major theme which Axelrod stresses throughout is that a strategy like TIT FOR TAT can emerge and be successful *regardless* of whether the participants exercise conscious foresight or not. It can arise simply out of the logic of the situation. Thus, in biological systems we do not need to suppose that organisms have fore-

sight in order to analyse how they can evolve diverse behavioural strategies.

HOW TO COOPERATE

The final chapters of the book are devoted to an analysis of the sociology of cooperation based on the foregoing theoretical investigations. Without examining all the intricacies of this analysis it is worth summarising the major points.

Axelrod provides four suggestions for doing well in an iterated Prisoner's Dilemma.

(1) *Don't be envious of the other player.* Because the Prisoner's Dilemma is a non-zero-sum game both sides can do well. In each encounter each player should concentrate on doing well for himself rather than on trying to do better than the other player. In the computer tournaments TIT FOR TAT did well by getting a higher overall score than any other player, even though it got less points than its opponents in each individual encounter.

(2) *Don't be the first to defect.* "Nasty" strategies are defined as those which are the first to defect. On average, nasty strategies did worse than "nice" ones (like TIT FOR TAT).

(3) *Reciprocate both cooperation and defection.* If the other player defects you must punish him by defecting yourself. But you shouldn't bear grudges. If the other player subsequently cooperates after defecting then you should cooperate again.

(4) *Don't be too clever.* In the computer tournament the sophisticated rules did not do better than the simple ones. This is in contrast to the situation in zero-sum games like chess where the more sophisticated a strategy the better it will be. One reason why, in the Prisoner's Dilemma, the sophisticated strategies were not particularly good is that their complexity made it more difficult for the other player to recognise what rules were being adopted. Consequently, it was more difficult to get cooperative patterns of behaviour established.

The iterated Prisoner's Dilemma requires that certain technical conditions must be met in order that a cooperative strategy like TIT FOR TAT can be successful. In biological systems, where there may be little or no conscious foresight, cooperation can emerge through trial and error, assuming the preconditions for cooperation are present. However, with *both* conscious foresight *and* the ability to shape our environment, human beings can actively promote cooperation, as opposed to just relying on trial and error. Axelrod gives five broad ways in which we can do this.

(1) *Enlarge the shadow of the future.* In other words, arrange things so that possible future interactions are sufficiently important.

(2) *Change the payoffs.* Remember that the Prisoner's Dilemma depends on Unilateral Defection > Mutual Cooperation > Mutual Defection > Unilateral Cooperation. It is also necessary that Mutual Cooperation (Unilateral Defection + Unilateral Cooperation)/2. This last requirement is not obvious but we need not explain it here.

The idea behind changing the payoffs is that, where individuals do not have private incentives to cooperate but it is socially useful that they do so, we can alter the incentives by, say, passing laws. For example, laws can be passed forcing people to honour their contracts or suffer the consequences. These consequences, fines or imprisonment, are such that the payoffs for not cooperating are not as attractive as would be the case if the laws were absent. Thus cooperation is induced.

(3) *Teach people to care about each other.* Cooperation will obviously be easier if we care about each other to some extent. "Caring" does not necessarily have to imply actively helping others. It may mean simply being concerned with not harming people (by selling them faulty goods, for example.)

(4) *Teach reciprocity.*

(5) *Improve recognition abilities.* You've got to be able to recognise the other player from past interactions so that you can practice reciprocity and encourage cooperation in the other player. Allied to this is the necessity to be able to recognise what the other player has done (or is doing). Lower organisms such as bacteria have a very limited ability to recognise other organisms so they tend to practice cooperation only where they have an exclusive relationship with one player (the host). Then they can attribute any changes in their environment to that one player. An example of poor recognition abilities hampering cooperation in a human social environment is that of superpower arms negotiations. In this case, the difficulty is more to do with recognising what the other player has done rather than with failing to recognise the other player.

A PROFOUND BOOK

In conclusion, the most exciting feature about *The Evolution of Cooperation* is how the analysis of an extremely simple game, The Prisoner's Dilemma, can shed so much light on such diverse fields of study. Not only is game theory an extremely powerful tool for understanding and explaining observed behaviour and evolutionary trends in biological systems but it can also shed light on the dynamics of human social systems both in spontaneous orders, such as the market, and in imposed orders, such as politics.

A friend of mine once told me that, although the quantity of books now being generated by mankind is unimaginably vast, the number of books that have anything profound to say is relatively small. I would place *The Evolution of Cooperation* firmly in this elite category.