The Evolution of Cooperation

Robert Axelrod

1.

The Problem of Cooperation

Under what conditions will cooperation emerge in a world of egoists without central authority? This question has intrigued people for a long time. And for good reason. We all know that people are not angels, and that they tend to look after themselves and their own first. Yet we also know that cooperation does occur and that our civilization is based upon it. But, in situations where each individual has an incentive to be selfish, how can cooperation ever develop?

The answer each of us gives to this question has a fundamental effect on how we think and act in our social, political, and economic relations with others. And the answers that others give have a great effect on how ready they will be to cooperate with us.

The most famous answer was given over three hundred years ago by Thomas Hobbes. It was pessimistic. He argued that before governments existed, the state of nature was dominated by the problem of selfish individuals who competed on such ruthless terms that life was “solitary, poor, nasty, brutish, and short” (Hobbes 1651/1962, p. 100). In his view, cooperation could not develop without a central authority, and consequently a strong government was necessary. Ever since, arguments about the proper scope of government have often focused on whether one could, or could not, expect cooperation to emerge in a particular domain if there were not an authority to police the situation.

Today nations interact without central authority. Therefore the requirements for the emergence of cooperation have relevance to many of the central issues of international politics. The most important problem is the security dilemma: nations often seek their own security through means which challenge the security of others. This problem arises in such areas as escalation of local conflicts and arms races. Related problems occur in international relations in the form of competition within alliances, tariff negotiations, and communal conflict in places like Cyprus.

The Soviet invasion of Afghanistan in 1979 presented the United States with a typical dilemma of choice. If the United States continued business as usual, the Soviet Union might be encouraged to try other forms of noncooperative behavior later on. On the other hand, any substantial lessening of United States cooperation risked some form of retaliation, which could then set off counter-retaliation, setting up a pattern of mutual hostility that could be difficult to end. Much of the domestic debate about foreign policy is concerned with problems of just this type. And properly so, since these are hard choices.

In everyday life, we may ask ourselves how many times we will invite acquaintances for dinner if they never invite us over in return. An executive in an organization does favors for another executive in order to get favors in exchange. A journalist who has received a leaked news story gives favorable coverage to the source in the hope that further leaks will be forthcoming. A business firm in an industry with only one other major company charges high prices with the expectation that the other firm will also maintain high prices—to their mutual advantage and at the expense of the consumer.

For me, a typical case of the emergence of cooperation is the development of patterns of behavior in a legislative body such as the United States Senate. Each senator has an incentive to appear effective to his or her constituents, even at the expense of conflicting with other senators who are trying to appear effective to *their* constituents. But this is hardly a situation of completely opposing interests, a zero-sum game. On the contrary, there are many opportunities for mutually rewarding activities by two senators. These mutually rewarding actions have led to the creation of an elaborate set of norms, or folkways, in the Senate. Among the most important of these is the norm of reciprocity—a folkway which involves helping out a colleague and getting repaid in kind. It includes vote trading but extends to so many types of mutually rewarding behavior that "it is not an exaggeration to say that reciprocity is a way of life in the Senate" (Matthews 1960, p. 100; see also Mayhew 1975).

Washington was not always like this. Early observers saw the members of the Washington community as quite unscrupulous, unreliable, and characterized by "falsehood, deceit, treachery" (Smith 1906, p. 190). In the 1980s the practice of reciprocity is well established. Even the significant changes in the Senate over the last two decades, tending toward more decentralization, more openness, and more equal distribution of power, have come without abating the folkway of reciprocity (Ornstein, Peabody, and Rhode 1977). As will be seen, it is *not* necessary to assume that senators are more honest, more generous, or more public-spirited than in earlier years to explain how cooperation based on reciprocity has emerged of proved stable. The emergence of cooperation can be explained as a consequence of individual senators pursuing their own interests.

The approach of this book is to investigate how individuals pursuing their own interests will act, followed by what effects this will have for the system as a whole. Put another way, the approach is to make some assumptions about individual motives and then deduce consequences for the behavior of the entire system (Schelling 1978). The case of the U.S. Senate if a good example, but the same style of reasoning can be applied to other settings.

The object of this enterprise is to develop a theory of cooperation that can be used to discover what is necessary for cooperation to emerge. By understanding the conditions that allow it to emerge, appropriate actions can be taken to foster the development of cooperation in a specific setting.

The Cooperation Theory that is presented in this book is based upon an investigation of individuals who pursue their own self-interest without the aid of a central authority to force them to cooperate with each other. The reason for assuming self-interest is that it allows an examination of the difficult case in which cooperation is not completely based upon a concern for others or upon the welfare of the group as a whole. It must, however, be stressed that this assumption is actually much less restrictive than it appears. If a sister is concerned for the welfare of her brother, the sister's self-interest can be thought of as including (among many other things) this concern for the welfare of her brother. But this does not necessarily eliminate all potential for conflict between sister and brother. Likewise, a nation may act in part out of regard for the interests of its friends, but this regard does not mean that even friendly countries are always able to cooperate for their mutual benefit. So the assumption of self-interest is really just an assumption that concern for others does not completely solve the problem of when to cooperate with them and when not to.

A good example of the fundamental problem of cooperation is the case where two industrial nations have erected trade barriers to each other's exports. Because of the mutual advantages of free trade, both countries would be better off if these barriers were eliminated. But if either country were to unilaterally eliminate its barriers, it would find itself facing terms of trade that hurt its own economy. In fact, whatever one country does, the other country is better off retaining its own trade barriers. Therefore, the problem is that each country has an incentive to retain trade barriers, leading to a worse outcome than would have been possible had both countries cooperated with each other.

This basic problem occurs when the pursuit of self-interest by each leads to a poor outcome for all. To make headway in understanding the vast array of specific situations which have this property, a way is needed to represent what is common to these situations without becoming bogged down in the details unique to each. Fortunately, there is such a representation available: the famous *Prisoner's Dilemma* game.

In the Prisoner's Dilemma game, there are two players. Each has two choices, namely cooperate or defect. Each must make the choice without knowing what the other will do. No matter what the other does, defection yields a higher payoff than cooperation. The dilemma is that if both defect, both do worse than if both had cooperated. This simple game will provide the basis for the entire analysis used in this book.

The way the game works is shown in figure 1. One player chooses a row, either cooperating or defecting. The other player simultaneously chooses a column, either cooperating or defecting. Together, these choices result in one of the four possible outcomes shown in that matrix. If both players cooperate, both do fairly well. Both get *R*, the *reward for mutual cooperation*. In the concrete illustration of figure 1, the reward is 3 points. This number might, for example, be a payoff in dollars that each player gets for that outcome. If one player cooperates but the other defects, the defecting player gets the *temptation to defect*, while the cooperating player gets the *sucker's payoff*. In the example, these are 5 points and 0 points respectively. If both defect, both get 1 point, the *punishment for mutual defection*.

What should you do in such a game? Suppose you are the row player, and you think the column player will cooperate. This means that you will get one of the two outcomes in the first column of figure 1. You have a choice. You can cooperate as well, getting the 3 points of the reward for mutual cooperation. Or you can defect, getting the 5 points of the temptation payoff. So it pays to defect if you think the other player will cooperate. But now suppose that you think the other player will defect. Now you are in the second column of figure 1, and you have a choice between cooperating, which would make you a sucker and give you 0 points, and defecting, which would result in mutual punishment, giving you 1 point. So it pays to defect if you think the other player will defect. This means that it is better to defect if you think the other player will cooperate, *and* it is better to defect if you think the other player will defect. So no matter what the other player does, it pays for you to defect.

So far, so good. But the same logic holds for the other player too. Therefore, the other player should defect no matter what you are expected to do. So you should both defect. But then you both get 1 point which is worse than the 3 points of the reward that you both could have gotten had you both cooperated. Individual rationality leads to a worse outcome for both than is possible. Hence the dilemma.

The Prisoner's Dilemma is simply an abstract formulation of some very common and very interesting situations in which what is best for each person individually leads to mutual defection, whereas everyone would have been better off with mutual cooperation. The definition of Prisoner's Dilemma requires that several relationships hold among the four different potential outcomes. The first relationship specifies the order of the four payoffs. The best a player can do is get *T*, the temptation to defect when the other player cooperates. The worst a player can do is get *S*, the sucker's payoff for cooperating while the other player defects. In ordering the other two outcomes, *R*, the reward for mutual cooperation, is assumed to be better than *P*, the punishment for mutual defection. This leads to a preference ranking of the four payoffs from best to worst as *T*, *R*, *P*, and *S*.

The second part of the definition of the Prisoner's Dilemma is that the players cannot get out of their dilemma by taking turns exploiting each other. This assumption means that an even chance of exploitation and being exploited is not as good an outcome for a player as mutual cooperation. It is therefore assumed that the reward for mutual cooperation is greater than the average of the temptation and the sucker's payoff. This assumption, together with the rank ordering of the four payoffs, defines the Prisoner's Dilemma.

Thus two egoists playing the game *once* will both choose their dominant choice, defection, and each will get less than they both could have gotten if they had cooperated. If the game is played a known finite number of times, the players still have no incentive to cooperate. This is certainly true on the last move since there is no future to influence. On the next-to-last move neither player will have an incentive to cooperate since they can both anticipate a defection by the other player on the very last move. Such a line of reasoning implies that the game will unravel all the way back to mutual defection on the first move of any sequence of plays that is of known finite length (Luce and Raiffa 1957, pp. 94–102). This reasoning does not apply if the players will interact an indefinite number of times. And in most realistic settings, the players cannot be sure when the last interaction between them will take place. As will be shown later, with an indefinite number of interactions, cooperation can emerge. The issue then becomes the discovery of the precise conditions that are necessary and sufficient for cooperation to emerge.

[…] A variety of ways to resolve the Prisoner's Dilemma have been developed. Each involves allowing some additional activity that alters the strategic interaction in such a way as to fundamentally change the nature of the problem. The original problem remains, however, because there are many situations in which these remedies are not available. Therefore, the problem will be considered in its fundamental form, without these alterations.

1. There is no mechanism available to the players to make enforceable threats or commitments (Schelling 1960). Since the players cannot commit themselves to a particular strategy, each must take into account all possible strategies that might be used by the other player. Moreover the players have all possible strategies available to themselves.
2. There is no way to be sure what the other player will do on a given move. This eliminates the possibility of metagame analysis (Howard 1971) which allows such options as "make the same choice as the other is about to make." It also eliminates the possibility of reliable reputations such as might be based on watching the other player interact with third parties. Thus the only information available to the players is the history of their interaction so far.
3. There is no way to eliminate the other player or run away from the interaction. Therefore each player retains the ability to cooperate or defect on each move.
4. There is no way to change the other player's payoffs. They payoffs already include whatever consideration each player has for the interests of the other (Taylor 1976, pp. 69–73).

Under these conditions, words not backed by actions are so cheap as to become meaningless. The players can communicate with each other only through the sequence of their own behavior. This is the problem of the Prisoner's Dilemma in its fundamental form.

What makes it possible for cooperation to emerge is the fact that the players might meet again. This possibility means that the choices made today not only determine the outcome of this move, but can also influence the later choices of the players. The future can therefore cast a shadow back upon the present and thereby affect the current strategic situation.

But the future is less important than the present—for two reasons. The first is that players tend to value payoffs less as the time of their obtainment recedes into the future. The second is that there is always some chance that the players will not meet again. An ongoing relationship may end when one or the other players moves away, changes jobs, dies, or goes bankrupt.

For these reasons, the payoff of the next move always counts less than the payoff of the current move. A natural way to take this into account is to cumulate payoffs over time in such a way that the next move is worth some fraction of the current move (Shubik 1970). The *weight* (or importance) of the next move relative to the current move will be called *w*. It represents the degree to which the payoff of the each move is discounted relative to the previous move, and is therefore a *discount parameter*.

[…] The first question you are tempted to ask is, "What is the best strategy?" In other words, what strategy will yield a player the highest possible score? This is a good question, but as will be shown later, no best rule exists independently of the strategy being used by the other player. In this sense, the iterated Prisoner's Dilemma is completely different from a game like chess. A chess master can safely use the assumption that the other player will make the most feared move. This assumption provides a basis for planning in a game like chess, where the interests of the players are completely antagonistic. But the situations represented by the Prisoner's Dilemma game are quite different. The interests of the players are not in total conflict. Both players can do well by getting the reward, *R*, for mutual cooperation, or both can do poorly by getting the punishment, *P*, for mutual defection. Using the assumption that the other player will always make the move you fear most will lead you to expect that the other will never cooperate, which in turn will lead you to defect, causing unending punishment. So unlike chess, in the Prisoner's Dilemma it is not safe to assume that the other player is out to get you.

In fact, the strategy that works best depends directly on what strategy the other player is using and, in particular, on whether this strategy leaves room for the development of mutual cooperation. This principle is based on the weight of the next move relative to the current move being sufficiently large to make the future important. In other words, the discount parameter, *w*, must be large enough to make the future loom large in the calculation of total payoffs. After all, if you are unlikely to meet the other person again, or if you care little about future payoffs, then you might as well defect now and not worry about the consequences for the future.

This leads to the first formal proposition. It is the sad news that if the future is important, there is no one best strategy.

[…] However, saying that a continuing chance of interaction is necessary for the development of cooperation is not the same as saying that it is sufficient. The demonstration that there is not a single best strategy leaves open the question of what patterns of behavior can be expected to emerge when there actually is a sufficiently high probability of continuing interaction between two individuals.

Before going on to study the behavior that can be expected to emerge, it is a good idea to take a closer look at which features of reality the Prisoner's Dilemma framework is, and is not, able to encompass. Fortunately, the very simplicity of the framework makes it possible to avoid many restrictive assumptions that would otherwise limit the analysis:

1. The payoffs of the players need not be comparable at all. For example, a journalist might get rewarded with another inside story, while the cooperating bureaucrat might be rewarded with a chance to have a policy argument presented in a favorable light.
2. The payoffs certainly do not have to be symmetric. It is a convenience to think of the interaction as exactly equivalent from the perspective of the two players, but this is not necessary. One does not have to assume, for example, that the reward for mutual cooperation, or any of the other three payoff parameters, have the same magnitude for both players. As mentioned earlier, one does not

even have to assume that they are measured in comparable units. The only thing that has to be assumed is that, for each player, the four payoffs are ordered as required for the definition of the Prisoner's Dilemma.

3. The payoffs of a player do not have to be measured on an absolute scale. They need only be measured relative to each other.
4. Cooperation need not be considered desirable from the point of view of the rest of the world. There are times when one wants to retard, rather than foster, cooperation between players. Collusive business practices are good for the businesses involved but not so good for the rest of society. In fact, most forms of corruption are welcome instances of cooperation for the participants but are unwelcome to everyone else. So, on occasion, the theory will be used in reverse to show how to prevent, rather than to promote, cooperation.
5. There is no need to assume that the players are rational. They need not be trying to maximize their rewards. Their strategies may simply reflect standard operating procedures, rules of thumb, instincts, habits, or imitation.
6. The actions that players take are not necessarily even conscious choices. A person who sometimes returns a favor, and sometimes does not, may not think about what strategy is being used. There is no need to assume deliberate choice at all.

The framework is broad enough to encompass not only people but also nations and bacteria. Nations certainly take actions which can be interpreted as choices in a Prisoner's Dilemma—as in the raising or lowering of tariffs. It is not necessary to assume that such actions are rational or are the outcome of a unified actor pursuing a single goal. On the contrary, they might well be the result of an incredibly complex bureaucratic politics involving complicated information processing and shifting political coalitions.

Likewise, at the other extreme, an organism does not need a brain to play a game. Bacteria, for example, are highly responsive to selected aspects of their chemical environment. They can therefore respond differentially to what other organisms are doing, and these conditional strategies of behavior can be inherited. Moreover, the behavior of a bacterium can affect the fitness of other organisms around it, just as the behavior of other organisms can affect the fitness of a bacterium.

[…] Of course, the abstract formulation of the problem of cooperation as a Prisoner's Dilemma puts aside many vital features that make any actual interaction unique. Examples of what is left out by this formal abstraction include the possibility of verbal communication, the direct influence of third parties, the problems of implementing a choice, and the uncertainty about what the other player actually did on the preceding move. It is clear that the list of potentially relevant factors that have been left out could be extended almost indefinitely. Certainly no intelligent person should make an important choice without trying to take such complicating factors into account. The value of an analysis without them is that it can help to clarify some of the subtle features of the interaction—features which might otherwise be lost in the maze of complexities of the highly particular circumstances in which choice must actually be made. It is the very complexity of reality which makes the analysis of an abstract interaction so helpful as an aid to understanding.

*

Chapter excerpted from *The Evolution of Cooperation* by Robert Axelrod, originally published in 1984, new edition 2006.

From the book jacket:

Robert Axelrod is Professor of Political Science and Public Policy at the University of Michigan. A MacArthur Prize Fellow, he is a leading expert on game theory, artificial intelligence, evolutionary biology, mathematical modeling, and complexity theory. He lives in Ann Arbor, Michigan.