The Evolution of Cooperation*

Robert Axelrod

Professor of Political Science and Public Policy, University of Michigan, Ann Arbor. Dr. Axelrod is a member of the American National Academy of Sciences and the American Academy of Arts and Sciences. His honors include a MacArthur Foundation Fellowship for the period 1987 through 1992.

Under what conditions will cooperation emerge in a world of egoists without central authority? This question has intrigued people for a long time. 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.

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 eliminate its barriers unilaterally, 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.

---

* Adapted from Robert Axelrod, *The Evolution of Cooperation.* New York: Basic Books, 1984. Reprinted by permission.

The Computer Tournament

This basic problem occurs when the pursuit of self-interest by each leads to a poor outcome for all. To understand the vast array of specific situations like this, we need a way to represent what is common to them without becoming bogged down in the details unique to each. Fortunately, there is such representation available: the famous Prisoner's Dilemma game, invented about 1950 by two Rand Corporation scientists. In this game there are two players. Each has two choices, namely "cooperate" or "defect." The game is called the Prisoner's Dilemma because in its original form two prisoners face the choice of informing on each other (defecting) or remaining silent (cooperating). Each must make the choice without knowing what the other will do. One form of the game pays off as follows:

Player's Choice	*Payoff*
If both players defect:	Both players get $1.
If both players cooperate:	Both players get $3.
If one player defects while the other player cooperates:	The defector gets $5 and the cooperator gets zero.

One can see that no matter what the other player does, defection yields a higher payoff than cooperation. If you think the other player will cooperate, it pays for you to defect (getting $5 rather than $3). On the other hand, if you think the other player will defect, it still pays for you to defect (getting $1 rather than zero). Therefore the temptation is to defect. But, the dilemma is that if both defect, both do worse than if both had cooperated.

To find a good strategy to use in such situations, I invited experts in game theory to submit programs for a computer Prisoner's Dilemma tournament – much like a computer chess tournament. Each of these strategies was paired off with each of the others to see which would do best overall in repeated interactions.

Amazingly enough, the winner was the simplest of all candidates submitted. This was a strategy of simple reciprocity which cooperates on the first move and then does whatever the other player did on the previous move. Using an American colloquial phrase, this strategy was named Tit for Tat. A second round of the tournament was conducted in which many more entries were submitted by amateurs and professionals alike, all of whom were aware of the results of the first round. The result was another victory for simple reciprocity.

The analysis of the data from these tournaments reveals four properties which tend to make a strategy successful: avoidance of unnecessary conflict by cooperating as long as the other player does, provocability in the

face of an uncalled-for defection by the other, forgiveness after responding to a provocation, and clarity of behavior so that the other player can recognize and adapt to your pattern of action.

"The soldiers of these opposing small units actually violated orders from their own high commands in order to achieve tacit cooperation with each other… cooperation based upon reciprocity can develop even between antagonists."

Live and Let Live in World War I

One concrete demonstration of this theory in the real world is the fascinating case of the "live and let live" system that emerged during the trench warfare of the western front in World War I. In the midst of this bitter conflict, the frontline soldiers often refrained from shooting to kill – provided their restraint was reciprocated by the soldiers on the other side.

For example, in the summer of 1915, a soldier saw that the enemy would be likely to reciprocate cooperation based on the desire for fresh rations.

> It would be child's play to shell the road behind the enemy's trenches, crowded as it must be with ration wagons and water carts, into a bloodstained wilderness … but on the whole there is silence. After all, if you prevent your enemy from drawing his rations, his remedy is simple: He will prevent you from drawing yours. (1)

> In one section the hour of 8 to 9 a.m. was regarded as consecrated to "private business," and certain places indicated by a flag were regarded as out of bounds by the snipers on both sides. (2)

What made this mutual restraint possible was the static nature of trench warfare, where the same small units faced each other for extended periods of time. The soldiers of these opposing small units actually violated orders from their own high commands in order to achieve tacit cooperation with each other.

This case illustrates the point that cooperation can get started, evolve, and prove stable in situations which otherwise appear extraordinarily unpromising. In particular, the "live and let live" system demonstrates that friendship is hardly necessary for the development of cooperation. Under suitable conditions, cooperation based upon reciprocity can develop even between antagonists.

Conditions for Stable Cooperation

Much more can be said about the conditions necessary for cooperation to emerge, based on thousands of games in the two tournaments, theoretical proofs, and corroboration from many real-world examples. For instance, the individuals involved do not have to be rational: The evolutionary process allows successful strategies to thrive, even if the players do not know why or how. Nor do they have to exchange messages or commitments: They do not need words, because their deeds speak for them. Likewise, there is no need to assume trust between the players: The use of reciprocity can be enough to make defection unproductive. Altruism is not needed: Successful strategies can elicit cooperation even from an egoist. Finally, no central authority is needed: Cooperation based on reciprocity can be self-policing.

"For cooperation to prove stable, the future must have a sufficiently large shadow . . . the importance of the next encounter between the same two individuals must be great enough to make [noncooperation] an unprofitable strategy."

For cooperation to emerge, the interaction must extend over an indefinite (or at least an unknown) number of moves, based on the following logic: Two egoists playing the game once will both be tempted to choose defection since that action does better no matter what action the other player takes. If the game is played a known, finite number of times, the players likewise have no incentive to cooperate on the last move, nor on the next-to-last move since both can anticipate a defection by the other player. Similar reasoning implies that the game will unravel all the way back to mutual defection on the first move. It need not unravel, however, if the players interact an indefinite number of times. And in most settings, the players cannot be sure when the last interaction between them will take place. An indefinite number of interactions, therefore, is a condition under which cooperation can emerge.

For cooperation to prove stable, the future must have a sufficiently large shadow. This means that the importance of the next encounter between the same two individuals must be great enough to make defection an unprofitable strategy. It requires that the players have a large enough chance of meeting again and that they do not discount the significance of their next meeting too greatly. For example, what made cooperation possible in the trench warfare of World War I was the fact that the same small units from opposite sides of no-man's-land would be in contact for

long periods of time, so if one side broke the tacit understandings, then the other side could retaliate against the same unit.

In order for cooperation to get started in the first place, one more condition is required. The problem is that in a world of unconditional defection, a single individual who offers cooperation cannot prosper unless some others are around who will reciprocate. On the other hand, cooperation can emerge from small clusters of discriminating individuals as long as these individuals have even a small proportion of their interactions with each other. So there must be some clustering of individuals who use strategies with two properties: The strategy cooperates on the first move, and discriminates between those who respond to the cooperation and those who do not.

"Once the US and the USSR know that they will be dealing with each other indefinitely, the necessary preconditions for cooperation will exist. . . . The foundation of cooperation is not really trust, but the durability of the relationship."

If a so-called "nice" strategy (that is, one which is never the first to defect) does eventually come to be adopted by virtually everyone, then individuals using this nice strategy can afford to be generous in their opening moves with any others. In fact, a population of nice strategies can also protect itself from clusters of individuals using any other strategy just as well as it can protect itself against single individuals.

Evolution of Cooperation

The tournament results give a chronological picture of the evolution of cooperation. Cooperation can begin with small clusters. It can thrive with strategies that are "nice" (that is, never the first to defect), provocable, and somewhat forgiving. Once established in a population, individuals using such discriminating strategies can protect themselves from invasion. The overall level of cooperation tends to go up and not down. In other words, the machinery for the evolution of cooperation contains a "ratchet," that is, it increases. Many institutions have developed stable patterns of cooperation based upon similar norms. Diamond markets, for example, are famous for the way their members exchange millions of dollars worth of goods with only a verbal pledge and a handshake. The key factor is that the participants know they will be dealing with each other again and again. Therefore any attempt to exploit the situation will simply not pay.

In other contexts, mutually rewarding relations become so commonplace that the separate identities of the participants can become blurred. For example, Lloyd's of London began as a small group of independent insurance brokers. Since the insurance of a ship and its cargo would be a large undertaking for one dealer, several brokers frequently made trades with each other to pool their risks. The frequency of the interactions was so great that the underwriters gradually developed into a federated organization with a formal structure of its own. The potential for attaining cooperation without formal agreements has its bright side in other contexts. For example, it means that cooperation on the control of the arms race does not have to be sought entirely through the formal mechanism of negotiated treaties. Arms control could also evolve tacitly. Once the US and the USSR know that they will be dealing with each other indefinitely, the necessary preconditions for cooperation will exist. The leaders may not like each other, but neither did the soldiers in World War I who learned to live and let live.

The foundation of cooperation is not really trust, but the durability of the relationship. When the conditions are right, the players can come to cooperate with each other through trial-and-error learning about possibilities for mutual rewards, through imitation of other successful players, or even through a blind process of selection of the more successful strategies with a weeding out of the less successful ones. Whether the players trust each other or not is less important in the long run than whether the conditions are ripe for them to build a stable pattern of cooperation with each other.

The Value of Provocability

Cooperation theory has implications for individual choice as well as for the design of institutions. Speaking personally, one of my biggest surprises in working on this project has been the value of provocability and that it is important to respond sooner, rather than later. I came to this project believing one should be slow to anger. The results of the computer tournament for the Prisoner's Dilemma demonstrate that it is actually better to respond quickly to a provocation. It turns out that if one waits to respond to uncalled-for defections, there is a risk of sending the wrong signal. The longer defections are allowed to go unchallenged, the more likely it is that the other player will draw the conclusion that defection can pay. And the more strongly this pattern is established, the harder it will be to break it. The success of simple reciprocity certainly illustrates this point. By responding right away, it gives the quickest possible feedback that a defection will not pay.

The response to potential violations of arms control agreements illustrates this point. Each superpower has occasionally taken steps which appear to be designed to probe the limits of its agreements with the other. The sooner the other detects and responds (in moderation) to these probes, the better. Waiting for probes to accumulate only risks the need for a response so large as to evoke yet more trouble.

The speed of response depends upon the time required to detect a given choice by the other player. The shorter this time is, the more stable cooperation can be. A rapid detection means that the next move in the interaction comes quickly, thereby increasing the shadow of the future. For this reason, the only arms control agreements which can be stable are those whose violations can be detected soon enough. The critical requirement is that violations can be detected before they can accumulate to such an extent that the victim's provocability is no longer enough to prevent the challenger from having an incentive to defect.

A Self-Reinforcing Ratchet Effect

Once the word gets out that reciprocity works – among nations or among individuals - it becomes the thing to do. If you expect others to reciprocate your defections as well as your cooperations, you will be wise to avoid starting any trouble. Moreover, you will be wise to respond appropriately after someone else defects, showing that you will not be exploited. Thus you too would be wise to use a strategy based upon reciprocity. So would everyone else. In this manner the appreciation of the value of reciprocity becomes self-reinforcing. Once it gets going, it gets stronger and stronger.

". . . simple reciprocity succeeds without doing better than anyone with whom it interacts. It succeeds by eliciting cooperation from others, not by defeating them"

This is the essence of the ratchet effect: Once cooperation based upon reciprocity gets established in a population, it cannot be overcome even by a cluster of individuals who try to exploit the others. The establishment of stable cooperation can take a long time if it is based upon blind forces of evolution, or it can happen rather quickly if its operation can be appreciated by intelligent players. The empirical and theoretical results might help people see more clearly the opportunities for reciprocity latent in their world. Knowing the concepts that accounted for the results of the two rounds of the computer Prisoner's Dilemma tournament, and knowing the reasons and conditions for the success of reciprocity, might provide some additional foresight.

From National Competitiveness to Global Cooperation

Robert Gilpin points out that from the ancient Greeks to contemporary scholarship all political theory addressed one fundamental question: "How can the human race, whether for selfish or more cosmopolitan ends, understand and control the seemingly blind forces of history?" (3) In the contemporary world this question has become especially acute because of the development of nuclear weapons.

Today, the most important problems facing humanity are in the arena of international relations, where independent, egoistic nations face each other in a state of near anarchy. Many of these problems take the form of an iterated Prisoner's Dilemma. Examples can include arms races, nuclear proliferation, crisis bargaining, and military escalation.

Therefore, the advice to players of the Prisoner's Dilemma might serve as good advice to national leaders as well: Don't be envious, don't be the first to defect, reciprocate both cooperation and defection, and don't be too clever.

There is a lesson in the fact that simple reciprocity succeeds without doing better than anyone with whom it interacts. It succeeds by eliciting cooperation from others, not by defeating them. We are used to thinking about competitions in which there is only one winner, competitions such as football or chess. But the world is rarely like that. In a vast range of situations, mutual cooperation can be better for both sides than mutual defection. The key to doing well lies not in overcoming others, but in eliciting their cooperation.

References

1. Ian Hay, *The First Hundred Thousand* (London: Wm. Blackwood, 1916).

2. John H. Morgan, *Leaves from a Field Note-Book* (London: Macmillan, 1916).

3. Robert Gilpin, *War and Change in World Politics* (Cambridge: Cambridge University Press, 1981).