RECIPROCITY AND THE LAW

Christopher Badcock*

I. THE FREE-RIDER THEOREM

One thing that law has in common with modern evolutionary biology is that both are concerned with the problem of free-riders: *individuals who get the benefits of social cooperation without paying the costs*. This is especially true with respect to enforcement and corrections agencies. In law enforcement, most offenders are clearly free-riders in one sense or another. However, legal scholars will not be so familiar with the situation in evolutionary biology, so let me briefly explain how free-riders fit in there.

In evolution, the central difficulty in explaining social cooperation is the fact that natural selection comes about through competition between individuals for reproductive success. Those individuals whose heritable traits promote their reproductive success, by definition, leave more copies of the genes for those traits than those who do not. Since an organism's reproductive success is critically dependent on the resources it can acquire and control, it is easy to see how self-interested behavior can be explained. It is more difficult to see how cooperative or even self-sacrificing behavior might evolve. Yet cooperation and self-sacrifice are surprisingly common in nature.

Characteristically, the problem worried Darwin, who feared that natural selection might never be able to explain the existence of cooperation and self-sacrifice in nature. However, Darwin's followers complacently believed that natural selection could explain even the most extreme forms of self-sacrifice by acting on whole groups such as families, herds, populations or even the entire species. This was known as "group selection" and appeared to provide a solution by claiming that, if self-sacrifice benefitted the group or species, and if selection acted on the group or species, self-sacrifice would be selected.

Consider a practical example. Suppose there is a species whose members have evolved a gene that halves their reproductive success and thereby promotes the interests of the whole by limiting population growth and conserving food supplies. Such an adaptation is easy to imagine and could indeed be controlled by a single gene—for example, an enzyme involved in reproduction or appetite. Now suppose that a mutant appears which lacks the copy of the gene for the enzyme and doubles its reproductive success. Obviously, such a mutant will have twice as many descendants as its self-denying conspecifics, and before long the entire species will be made up of such mutants. The mutants would constitute a group of free-riders to the extent that they were exploiting the reproductive restraint of others in the

---

* Associate Professor, London School of Economics.

species to their own benefit. The food resources conserved by others would be used to feed their own offspring twice as often as it would be used to feed the members of the population with the gene that halved their reproductive success.

II. PRISONER'S DILEMMA

In order to account for social behavior, we need an approach that takes both cooperation with the group and selfish exploitation of it seriously, otherwise we will be only telling half the story. The smallest such group would be one of two people, and the following situation might represent the simplest form of social cooperation.

Suppose that you and an acquaintance go into a restaurant for a meal. Considered in the most basic terms, we might say that both you and your partner could obtain a benefit from such an event, and pay a cost. The benefit would be the meal, and the cost, the bill. Each of you would face a choice: either to pay or not to pay. Let's call paying the bill *cooperating*, and not paying it *defecting*. There are four possibilities, depending on whether you cooperate or defect, and whether your partner cooperates or defects. If you both cooperate, you both pay and you both eat. We could call this *reciprocity* because cost and benefit are equally shared.

Now suppose that you cooperate and pay, but your partner does not, perhaps by claiming not to have enough money or simply refusing to do so. If you pay for both meals, we could call this *altruism* on your part in the sense that you paid for your partner's lunch. However, from your point of view we could claim that your partner had acted *selfishly* because he got a free lunch.

Finally, imagine a situation where neither of you were prepared to pay anything. The result, of course, is that neither of you got any lunch. This we might call *spite*. The spiteful aspect would be obvious if, for example, you refused to countenance eating out with your partner once he indicated that he was not prepared to pay.

If we now think about the relative value of these outcomes to you or your partner, we can see that a free lunch is always best because you get a benefit without a cost. This is better than a lunch with a shared cost, because, although you still get your lunch, you must pay for it. This, in turn, is better than no lunch at all. Although it imposes no cost, it also carries no benefit whatsoever. Finally, the worst outcome is having to pay for someone else's lunch; a cost without a benefit to you. Translating these examples back into the terms we defined in the previous paragraphs, we can see that selfishness (a free lunch) is better than reciprocity (both eating, both paying), but that, in turn, is worth more than spite (neither pays, neither eats). The worst outcome is having to perform an act of altruism (you pay, the other eats at your

expense). This example is usually called a "Prisoner's Dilemma," stemming from the common practice of separately interrogating suspects who may have collaborated in a crime and offering inducements to each to incriminate the other.

Prisoner's Dilemma is uninteresting if limited to a so-called single-shot interaction. Clearly, if two parties only interact once, it will always pay for each to incriminate the other, or defect. In this way, they defend themselves from the worst outcome—cooperating when the other defects—but stand to win if the other cooperates. However, things can be very different if Prisoner's Dilemma-style interactions between the same two players are repeated over a period of time, or in the technical term, *iterated.*

For example, some years ago my wife and another woman made an agreement to cooperate in taking their children to school. One mother agreed to take the children to school at the beginning of the day and the other agreed to collect them at the end. Obviously, both had a real incentive to cooperate in this way because it meant halving the time spent taking and collecting, petrol consumed and so on. At first, all went well, but towards the end of the first term, the other lady's car began to let her down and my wife made significantly more journeys than she did. The vacations came and went, and the next term began on equitable footing again. However, as the end of the term approached, the other lady's car once again began to become afflicted with mechanical problems. Finally, when exactly the same thing occurred at the end of the third term, my wife made a polite excuse and established another, more successful, arrangement with another mother for the next school year.

III. COMPUTING COOPERATION

From time to time, new publications appear in fields of science which instantly make everything previously published on the same subject obsolete. An obvious example would be Charles Darwin's *Origin of Species,* but a more recent one is the series of publications culminating in Robert Axelrod's book, *The Evolution of Cooperation.*[1] At first sight, nothing could be more surprising because the book's ostensible purpose was to report and comment on nothing more significant than the results of two computer tournaments. Yet the consequences of how and why the winning program won ramify throughout social and biological science, making Axelrod's book a landmark publication in the modern study of social behavior.

---

1. R. AXELROD, THE EVOLUTION OF COOPERATION 137 (1984).

Axelrod invited participants to submit computer programs to play iterated Prisoner's Dilemma. We have already met a real-life example of this at the end of the last section where we saw that, effectively, my wife was playing a game of Prisoner's Dilemma when she participated in the arrangement to share the school journeys. As we have also already seen, Prisoner's Dilemma boils down to a simple choice of decisions, either to cooperate or to defect. The computer programs entered in Axelrod's tournaments each had to make a similar choice with only the knowledge of the previous choices of its opponent. Programs were played against each other in a round robin tournament (meaning that each entry was played against every other entry), and scores were totaled.

Two tournaments were played. In the first fourteen entries, individuals with Prisoner's Dilemma experience from five disciplines: psychology, economics, political science, mathematics and sociology submitted entries. The winner was "Tit-for-Tat," the shortest, simplest program submitted. Its strategy was to cooperate on the first move and thereafter do exactly what its opponent had done on the previous move. Tit-for-Tat won, despite the fact that it was known beforehand to be a good strategy, and a number of the programs submitted were attempted improvements upon it. Analysis showed that neither the discipline of the author, the length of the program, nor its brevity, could account for the result. In the first tournament, what distinguished higher from lower scoring programs was that the former were the "nice" ones—meaning they did not defect first. The "nice" rules did well in the tournament largely by virtue of their interactions with one another, and because there were enough of them to raise their individual scores.

Following the first tournament, the results and their analysis were circulated widely and a second tournament held. This time, sixty-two entries were submitted from six countries and contestants ranged from a ten-year-old computer hobbyist to professors of computing, biology and physics. Despite the success of Tit-for-Tat in the first round and the fact that entrants could submit any program, even one authored by someone else and entered by others in the first tournament, only Anatol Rapoport, who entered Tit-for-Tat the first time, did so again. The sixty-three programs were paired in 3969 different ways and over a million rounds were played. Once again, Tit-for-Tat won.

In a variation of the basic round robin tournament, an "evolutionary" or "ecological" variant was played, made possible, like the main tournament, only by the speed of modern computers. In this tournament, programs were "selected out" as they fell to the bottom of the score sheet, and their representation was weighted according to their past success. Once again, Tit-for-Tat emerged as the winner, and went on winning by a steadily growing margin.

How can Tit-for-Tat's astonishing success in these tournaments be explained? Fundamentally, its success relies on three principal attributes.

First, Tit-for-Tat was *robust*, meaning that it could not be easily exploited and could hold its own against many different opponents. Here it contrasts with a strategy entered in the second tournament called "Tit-for-Two-Tats." Tit-for-Two-Tats is just like Tit-for-Tat except that it only retaliates with a single defection after two defections by its opponent. Although Tit-for-Two-Tats would have won the first tournament, had it been entered, it failed to do well in the second. This was due to the fact that other strategies had been entered which exploited its distinctive characteristic of forgiving single defections against it. Tit-for-Tat had no such weakness, and so did just as well in the second tournament as it had in the first.

Second, Tit-for-Tat, like the majority of winning programs in both tournaments was "nice," which means, as we have already seen, that it would never be the first to defect. To see how Tit-for-Tat could elicit cooperation by being initially cooperative itself, consider how it plays against the worst possible strategy of all, one which cooperates unconditionally—what we might call "Sucker." Against its opposite, "Defector," which defects every time, Sucker cannot score any points at all in any number of games against Defector which gets winning points each time. However, if Sucker meets Tit-for-Tat, both score the pay-off for mutual cooperation on the first round and continue to do so indefinitely.

Finally, and perhaps most importantly, Tit-for-Tat was ideally *reciprocating*. In other words, it immediately and unfailingly rewarded cooperation with cooperation, and defection with defection. The latter is the secret of its robustness, the former of its "niceness." This also gave it clarity and predictability. Its decision rule was clear and simple, building confidence in its reliability. Astonishingly, its success was based on the fact that it never won any more points than any competitor it played against. "It won," as Axelrod points out, "not by doing better than any other player, but by eliciting cooperation from the other player. In this way Tit-for-Tat does well by promoting the mutual interest rather than by exploiting the other's weakness. A moral person couldn't do much better."[2]

IV. THE REWARDS OF RECIPROCITY

At this point it is important to correct a common misunderstanding relating to the role of reciprocity in establishing and maintaining cooperation. This is the view that the success of Tit-for-Tat represents a return to Old

---

2. *Id.*

Testament-style "eye-for-an-eye" and "tooth-for-a-tooth" retaliation.[3] This approach embodies the notion of a blood-feud or vendetta which, while sometimes encouraging cooperation, also maintains violence, once started, in a self-perpetuating cycle.

A case which shows that reciprocity does not necessarily breed self-perpetuating cycles of violence is provided by one of the most dramatic real-life examples of Tit-for-Tat-style reciprocity in establishing cooperation. Historical accounts from World War I show that an astonishing "live-and-let live" cooperation emerged between the opposing armies. A British officer who discovered such cooperation in a French sector recounted that it was maintained because the French only fired if fired upon, but always returned two shots for every one shot fired at them. The fact that they retaliated with twice as much force shows that they were not following a crude eye-for-an-eye tactic. Establishing a surprising degree of peaceful cooperation suggests that it was not an example of blood-feud or vendetta mentality.

On the contrary, the French were doing exactly what Tit-for-Tat did in the tournaments: not merely hitting back for the sake of some primitive law of talion, but reciprocating—not for its own sake—but in order to bring about continuing cooperation. This is why the French responded to the Germans with twice as much force as the Germans directed towards them, but this is also why the French never shot first. They sought to induce peaceful cooperation in the enemy by punishing them for their belligerent defections. Here reciprocity did not exist as an end in itself, but as a means to an end: cooperation with the other side. Essentially, this was how Tit-for-Tat behaved in the tournaments. Its immediate and discriminating reciprocity was added to its other attribute; it was never the first to defect. Tit-for-Tat was not the Prisoner's Dilemma equivalent of primitive talion because it began by cooperating and succeeded by encouraging its partners to cooperate, just as the French soldiers did with their two-for-one retaliations and with their parallel reluctance to start any shooting.

The live-and-let-live system, whereby one side would only inflict damage on the other if the other punished it, relied in large part on the fact that trench warfare produced a situation where small, evenly matched units faced one another for considerable periods of time. In other words, it was an iterated Prisoner's Dilemma, comparable to my wife's situation in the shared school run described earlier. As Axelrod points out in his discussion, cooperation in this situation was equivalent to avoiding damaging the enemy, while defection meant shooting to kill. In the short run, it was always preferable to damage the enemy irrespective of what he was doing in these situations because

---

3. *See generally* EXODUS 22:24.

damage inflicted now could weaken him in future conflicts. Again, mutual cooperation was preferable to mutual defection since the latter implied that both sides took casualties. Taken together, we have the classic set of inequalities required for Prisoner's Dilemma in the sense that killing all the enemy was always better than mutual survival, which in turn was always better than mutual casualties. The latter is always preferable to the destruction of our side without damage to the enemy.

The consequence was an astonishing and exquisite example of how cooperation can emerge spontaneously and be maintained, motivated purely by the self-interest of the cooperating parties. Like the triumph of Tit-for-Tat in Axelrod's computer tournaments, the emergence of the live-and-let-live system suggests that cooperation can emerge spontaneously in a world of egoists, motivated solely by self-interest. Peaceable restraint, which in the World War I example originated on both sides during meal-times or bad weather, became extended to the point where open fraternization became possible and where, on occasions, soldiers on one side would actually apologize to the other for unintended breaks in the informal truce. Sniper or artillery fire, ostensibly intended to inflict damage on the other side, was in fact used as a means of maintaining the peace. Snipers demonstrated their skill by repeatedly shooting at the same place on a wall to bore a neat hole, while artillery showed off their accuracy by hitting non-military targets. These apparently ritualized expressions of the conflict served to maintain the overall pattern of cooperation because they showed that, if necessary, the other side could be provoked into telling retaliation.

This situation is truly astonishing because of the way in which it totally contradicts the widely-held view that social order can only emerge if it is imposed or policed in some way. The cooperating parties involved in the live-and-let-live system were not subject to any external authority enforcing the peace. They did not belong to the same society and did not share a common language or set of values. They had not been socialized to cooperate with their enemies, nor did their cultural backgrounds promote such fraternization. On the contrary, the cooperating parties came from different cultural backgrounds which were at the time dominated by appalling nationalistic jingoism. Most outstanding of all was the fact that the participants in this astonishing display of international cooperation were subject to their respective High Commands and military disciplines which were diametrically opposed to what they were doing. This demonstrates that, even where external authority attempts to impose one kind of social order, order of a quite different kind can nevertheless emerge spontaneously, naturally and with greater effect.

It is a sobering thought that the imposed social order so popular with conventional social thought was the very thing which worked against the live-

and-let-live system, and which eventually destroyed it. The institution of unexpected raids, often carried out by troops not familiar with the local conventions, helped undermine the implicit truces which had come to dominate long tracts of the front line. These sudden and vicious defections roused the enemy to retaliate in kind and undermined the trust and mutual confidence built up earlier; eventually live-and-let-live was replaced by "kill-or-be-killed."

V. ENCOURAGING COOPERATION

The live-and-let-live cooperation of World War I exemplifies the basic principles that apply generally wherever cooperation emerges spontaneously, and apply to law and law enforcement as well.

First, "the future must have a sufficiently large shadow."[4] In other words, we must make interactions on-going and iterated, rather than one-shot and isolated. As we saw, there is no incentive to cooperate in a one-shot interaction, but there is a real pay-off to cooperation in continuing interactions. In practical terms, this means avoiding anonymity and encouraging future interaction, for example, by involving the local community with local law-breakers.

Second, we must help actors find other cooperators with whom to cooperate. Since prisons and penitentiaries are by definition full of defectors, one could see this as an incentive to avoid incarceration wherever possible and to encourage erstwhile defectors to mix with more cooperative players. For example, one could involve the community in correction and rehabilitation programs.[5]

Third, cooperation will be encouraged if actors can protect themselves from defectors by defecting against them. Individuals must realize that the cooperation of others is contingent on their cooperation. The community cannot be expected to issue blank checks where cooperation is concerned, or to tolerate defectors who exploit the community's cooperation. It means treating everyone as a potential cooperator, assuming that all participants in the community share the dignity and freedom, both to cooperate and to defect.

The triumph of Tit-for-Tat in Axelrod's tournaments and the live-and-let-live cooperation of World War I both convey the essential optimism of modern insights into the evolution of cooperation. As Robert Axelrod observed, "[T]he main results of Cooperation Theory are encouraging. They show that cooperation can get started by even a small cluster of individuals

---

4. *Id.* at 174.

5. For example, note the contributions of the Vermont Corrections Department.

who are prepared to reciprocate cooperation, even in a world where no one else will cooperate."[6] What worked in a world war deserves at least a chance in the war against crime.

6. AXELROD, *supra* note 1, at 173.