The Prisoner's Dilemma

Peter Singer

Peter Singer is a utilitarian moral philosopher best known for his pioneering book Animal Liberation *(1975), but he has also written and edited many other books and essays in the field of ethics, primarily on applied ethics, such as animal rights, abortion, and famine relief. He taught philosophy for many years at Monash University in Victoria, Australia, moving to Princeton University in 1999. The following is excerpted from his book,* The Expanding Circle *(1981).*

In the cells of the Ruritanian secret police are two political prisoners. The police are trying to persuade them to confess to membership in an illegal opposition party. The prisoners know that if neither of them confesses, the police will not be able to make the charge stick, but they will be interrogated in the cells for another three months before the police give up and let them go. If one of them confesses, implicating the other, the one who confesses will be released immediately but the other will be sentenced to eight years in jail. If both of them confess, their helpfulness will be taken into account and they will get five years in jail. Since the prisoners are interrogated separately, neither can know if the other has confessed or not.

The dilemma is, of course, whether to confess. The point of the story is that circumstances have been so arranged that if either prisoner reasons from the point of view of self-interest, she will find it to her advantage to confess; whereas taking the interests of the two prisoners together, it is obviously in their interests if neither confesses. Thus the first prisoner's self-interested calculations go like this: "If the other prisoner confesses, it will be better for me if I have also confessed, for then I will get five years instead of eight; and if the other prisoner does not confess, it will still be better for me if I confess, for then I will be released immediately, instead of being interrogated for another three months. Since we are interrogated separately, whether the other prisoner confesses has nothing to do with whether I confess — our choices are entirely independent of each other. So whatever happens, it will be better for me if I confess." The second prisoner's self-interested reasoning will, of course, follow exactly the same route as the first prisoner's, and will come to the same conclusion. As a result, both prisoners, if self-interested, will confess, and both will spend the next five years in prison. There was a way for them both to be out in three months, but because they were locked into purely self-interested calculations, they could not take that route.

What would have to be changed in our assumptions about the prisoners to make it rational for them both to refuse to confess? One way of achieving this would be for the prisoners to make an agreement that would bind them both to silence. But how could each prisoner be confident that the other would keep the agreement? If one prisoner breaks the agreement, the other will be in prison for a long time, unable to punish the cheater in any way. So each prisoner will reason: "If the other one breaks the agreement, it will be better for me if I break it too; and if the other one keeps the agreement, I will still be better off if I break it. So I will break the agreement."

Without sanctions to back it up, an agreement is unable to bring two self-interested individuals to the outcome that is best for both of them, taking their interests together. What has to be changed to reach this result is the assumption that the prisoners are motivated by self-interest alone. If, for instance, they are altruistic to the extent of caring as much for the interests of their fellow prisoner as they care for their own interests, they will reason thus: "If the other prisoner does not confess it will be better for us both if I do not confess, for then between us we will be in prison for a total of six months, whereas if I do confess the total will be eight years; and if the other prisoner does confess it will still be better if I do not confess, for then the total served will be eight years, instead of ten. So whatever happens, taking our interests together, it will be better if I don't confess." A pair of altruistic prisoners will therefore come out of this situation better than a pair of self-interested prisoners, *even from the point of view of self-interest*.

Altruistic motivation is not the only way to achieve

a happier solution. Another possibility is that the prisoners are conscientious, regarding it as morally wrong to inform on a fellow prisoner; or if they are able to make an agreement, they might believe they have a duty to keep their promises. In either case, each will be able to rely on the other not confessing and they will be free in three months.

The Prisoner's Dilemma shows that, paradoxical as it may seem, we will sometimes be better off if we are not self-interested. Two or more people motivated by self-interest alone may not be able to promote their interests as well as they could if they were more altruistic or more conscientious.

The Prisoner's Dilemma explains why there could be an evolutionary advantage in being genuinely altruistic instead of making reciprocal exchanges on the basis of calculated self-interest. Prisons and confessions may not have played a substantial role in early human evolution, but other forms of cooperation surely did. Suppose two early humans are attacked by a saber tooth cat. If both flee, one will be picked off by the cat; if both stand their ground, there is a very good chance that they can fight the cat off; if one flees and the other stands and fights, the fugitive will escape and the fighter will be killed. Here the odds are sufficiently like those in the Prisoner's Dilemma to produce a similar result. From a self-interested point of view, if your partner flees your chances of survival are better if you flee too (you have a 50 percent chance rather than none at all) and if your partner stands and fights you still do better to run (you are sure of escape if you flee, whereas it is only probable, not certain, that together you and your partner can overcome the cat). So two purely self-interested early humans would flee, and one of them would die. Two early humans who cared for each other, however, would stand and fight, and most likely neither would die. Let us say, just to be able to put a figure on it, that two humans cooperating can defeat a saber tooth cat on nine out of every ten occasions and on the tenth occasion the cat kills one of them. Let us also say that when a saber tooth cat pursues two fleeing humans it always catches one of them, and which one it catches is entirely random, since differences in human running speed are negligible in comparison to the speed of the cat. Then one of a pair of purely self-interested humans would not, on average, last more than a single encounter with a saber tooth cat; but one of a pair of altruistic humans would on average survive ten such encounters.

If situations analogous to this imaginary saber tooth cat attack were common, early humans would do better hunting with altruistic comrades than with self-interested partners. Of course, an egoist who could find an altruist to go hunting with him would do better still; but altruists who could not detect — and refuse to assist — purely self-interested partners would be selected against. Evolution would therefore favor those who are genuinely altruistic to other genuine altruists, but are not altruistic to those who seek to take advantage of their altruism. We can add, again, that the same goal could be achieved if, instead of being altruistic, early humans were moved by something like a sense that it is wrong to desert a partner in the face of danger.