

MORE ON THE PRISONER'S DILEMMA.

By Sean Crawford, Department of Philosophy, The Open University

Covenants struck without the sword are but words. - Thomas Hobbes, *Leviathan*.

Imagine you are an anti-government radical, and you and your revolutionary partner have been captured by the secret police and charged with sedition. The prosecutor interrogates each of you separately and offers each of you the following deal:

If one of you defects by incriminating your partner, while the partner remains silent, then the defector will be convicted of a lesser crime and his sentence reduced to one year for providing information, while the silent confederate will be convicted of a more serious crime and given a four-year sentence for his refusal to provide information.

If you both stay silent, there will be insufficient evidence to convict either of you of the more serious crime, and you will each receive a sentence of two years for a lesser offence.

If you both defect by incriminating each other, you will both be convicted of the more serious crime, but given reduced sentences of three years for providing information.

In short, each of you has two choices: **COOPERATE** with your confederate by remaining silent, or **DEFECT** from the confederacy by incriminating your partner.

We thus have four possible outcomes:

1. You defect and your partner cooperates (1 year for you, 4 for him).
2. You both cooperate (2 years each).
3. You both defect (3 years each).
4. You cooperate, and your confederate defects (4 for you, 1 for him).

These four outcomes are represented in the “payoff matrix.” From a purely selfish point of view, the best outcome for you is 1, followed by 2, 3 and 4; for your confederate 4 is the best option, followed by 2, 3 and 1. Given this ordering of preferences, you should, as a rational, purely selfish individual, always defect no matter what your confederate does. For consider: he will either defect or cooperate. If he defects, you should too, to avoid the worst possible outcome for you. If he cooperates, you should defect, as you will get the smallest possible sentence, your

preferred outcome. Your confederate reasons the same way, of course. So it seems you should both defect. But if so, then you both get the third best outcome whereas if you had both cooperated you would both have your second best outcome. In the language of **Game Theory**, defecting is a **Dominant Strategy** in a game with this payoff matrix, that is, the strategy that it is always best to adopt regardless of the strategy of the other player.

But the best strategy leaves both of you worse off than if you had both cooperated and remained silent! By cooperating you would both be better off than you would if you act as independent agents. Cooperating will not, of course, get you your personal optimal outcome, a one year sentence, but it will get a better result than either of you can achieve if you both defect. This is the famous **Prisoner's Dilemma (PD)**. Is there any way out of this?

You might think that the way out of the dilemma is to agree before hand to cooperate. But this won't help, for how do you know you can **trust** your confederate to keep his part of the bargain? For the same reasoning as above applies to the question whether you should keep or renege on your promise to cooperate. The dominant strategy will again be to renege on your promise thus producing a worse outcome than keeping the promise!

The PD game was discovered by the game theorists Flood and Dresher around 1950 who were both working for the Rand corporation at the time. The Rand corporation was a private company created out of the US Air Force at the end of the Second World War whose aim was to investigate the possibility of intercontinental nuclear warfare, in particular the prospects of pre-emptive nuclear strikes. Flood had actually been trying to represent the sale of a used car as a game and he and Dresher had been running experiments to test how people behaved in an interaction that was repeated. The game was peculiar enough for Flood and Dresher to show it to the Princeton mathematician A. W. Tucker, who proposed the tale of the two prisoners to illustrate it. The Rand corporation had hired as consultants many mathematicians interested in game theory, the most famous of which is perhaps Jon von Neumann (1903-57) who was a brilliant Hungarian mathematician who made crucial mathematical contributions to the Manhattan Project for the development of atomic weapons as well as to the design of the modern computer. Some speculate that von Neuman is the model for the eponymous mad scientist in Stanley Kubrick's 1963 Film *Dr. Strangelove*. Game theory's connections with film don't end there, however. The hero of the latest Hollywood movie *A Beautiful Mind* is another famous game theorist, Jon Nash, who also worked for the Rand corporation. Nash is the inventor of the highly influential but controversial concept of the Nash Equilibrium, a central concept in game theory.

The PD isn't merely an entertaining game, as those in the Rand corporation were fully aware. Many social interactions exhibit the structure of the PD. Take a simple economic exchanges, for example. Since in many cases the payment and delivery of goods are not perfectly synchronised, the opportunity arises to cheat on the deal (either not to pay or not to send the goods). The notorious 'free rider' problem is an

example of a PD involving many individuals. Consider the bus systems in many European countries. They are paid for by what is almost an honour system of passengers paying their fares. The busses can only be kept running if enough people pay the fare, but not everyone need pay the fare to keep the busses running. The best outcome for each passenger individually is for him not to pay the fare and for enough other people to pay so that the busses may still run. But if everyone adopts this strategy, the buses will no longer run and everyone will have to take taxis, which is a worse situation for everyone than if everyone paid the fare.

It seems that the provision for any kind of public good suffers from the same problem: think of public television like the BBC, a clean environment, or national defense, which are things we all want but for which the dominant strategy for each individual will always be not to contribute to producing the good in question. For example, either enough people will pay their TV licenses to provide the service or they won't. If enough people pay their TV license then it's best for me not to pay (because I get it for free then). But if not enough people pay their license, and the probability that my paying it will make the difference is very low, then again it's in my interest not to pay. Obviously, if everyone thinks like this then there will be no BBC; a worse situation than if everyone paid their license.

Another vivid example concerns disarmament. Nations face a choice of arming or disarming. From a purely self-interested point of view, the best situation for a nation is to be armed and all other nations to be disarmed. Since it is still better to be armed if other nations are armed, this makes arming a dominant strategy, which produces the familiar result that the nations involved get a worse outcome than they would if they all disarmed.

One way to get the **assurance** you need to be able to **trust** someone to keep their part of the bargain is for there to be a mechanism by which sanctions are imposed on those who renege on agreements. Reverting to our PD, imagine that the revolutionary group that the prisoners belong to has a way of enforcing the agreement between them to cooperate: namely, they will kill the defector.

The philosopher **Thomas Hobbes** (1588-1679) thought that the only way out of social situations like this was to have some mechanism which ensured that individuals did not defect, some way of enforcing that agreements or contracts are kept by imposing sanctions on those who break them. For Hobbes it was the State that was the agency of enforcement, the 'sword' in the epigram. Without the existence of the government and its laws, courts, and police we would all be in a 'state of nature', thought Hobbes, essentially structured like a PD, which is a "constant state of war, of one with all" and hence, in his immortal words, "the life of man, solitary, poor, nasty, brutish, and short." The heavy fine imposed on those caught watching TV without a license or those caught riding busses for free are relatively trivial examples of the fear of the 'sword' producing the public goods. Can you think of more important ones? In the eyes of many influenced by Hobbes the need to escape the PD is an argument for the creation of a strong State with the power to enforce agreements.

But is the ‘sword’ the only way out of the PD? Perhaps not. For what if you knew you would be playing the game more than once with the same people? What if the social situations we find ourselves in are more like an *Iterated* or *Repeated* PD than a one-off PD, as many of them certainly seem to be? Perhaps the trust necessary for cooperation to get off the ground can evolve spontaneously during a series of PDs, especially if punitive behaviour by cooperators is adopted towards defectors. For defectors’ reputations will be ruined and so no one will trust them in the future and so they may very well not receive the benefits from cooperative interactions in the future. According to this line of thought, cooperation early on in a series of PDs makes possible not only the cooperative benefits in that situation but the cooperative benefits in future situations. So it may be in your self-interest to cooperate after all! But then escaping the PD does not require an activist State. Or perhaps it still does require one. Can you think why? Or perhaps this is entirely the wrong way of looking at it! Maybe people are, by their very nature, by and large both trusting and trustworthy. Indeed, could it be that a general altruistic concern for the well being of others provides sufficient reason for people to cooperate, to keep their agreements? Or is Hobbes’ more cynical view of human nature as egoistic more true to the facts? Play the repeated game and see what you think.

You can learn more about the PD and political philosophy in general by taking the Open University’s course in political philosophy, from which this version of the PD web game is drawn.