The Prisoners' Dilemma

Author
F. Heylighen

Date
Apr 13, 1995 (modified)
Nov 30, 1993 (created)

Cooperation is usually analysed in game theory by means of a non-zero-sum game called the "Prisoner's Dilemma" (Axelrod, 1984). The two players in the game can choose between two moves, either "cooperate" or "defect". The idea is that each player gains when both cooperate, but if only one of them cooperates, the other one, who defects, will gain more. If both defect, both lose (or gain very little) but not as much as the "cheated" cooperator whose cooperation is not returned. The whole game situation and its different outcomes can be summarized by table 1, where hypothetical "points" are given as an example of how the differences in result might be quantified.

Table 1: outcomes for actor A (in words, and in hypothetical "points") depending on the combination of A's action and B's action, in the "prisoner's dilemma" game situation. A similar scheme applies to the outcomes for B.

The game got its name from the following hypothetical situation: imagine two criminals arrested under the suspicion of having committed a crime together. However, the police does not have sufficient proof in order to have them convicted. The two prisoners are isolated from each other, and the police visit each of them and offer a deal: the one who offers evidence against the other one will be freed. If none of them accepts the offer, they are in fact cooperating against the police, and both of them will get only a small punishment because of lack of proof. They both gain. However, if one of them betrays the other one, by confessing to the police, the defector will gain more, since he is freed; the one who remained silent, on the other hand, will receive the full punishment, since he did not help the police, and there is sufficient proof. If both betray, both will be punished, but less severely than if they had refused to talk. The dilemma resides in the fact that each prisoner has a choice between only two options, but cannot make a good decision without knowing what the other one will do.

Such a distribution of losses and gains seems natural for many situations, since the cooperator whose action is not returned will lose resources to the defector, without either of them being able to collect the additional gain coming from the "synergy" of their cooperation. For simplicity we might consider the Prisoner's dilemma as zero-sum insofar as there is no mutual cooperation: either each gets 0 when both defect, or when one of them cooperates, the defector gets + 10, and the cooperator - 10, in total 0. On the other hand, if both cooperate the resulting synergy creates an additional gain that makes the sum positive: each of them gets 5, in total 10.

The gain for mutual cooperation (5) in the prisoner's dilemma is kept smaller than the gain for one-sided defection (10), so that there would always be a "temptation" to defect. This assumption is not generally valid. For example, it is easy to imagine that two wolves together would be able to kill an animal that is more than twice as large as the largest one each of them might have killed on his own. Even if an altruistic wolf would kill a rabbit and give it to another wolf, and the other wolf would do nothing in return, the selfish wolf would still have less to eat than if he had helped his companion to kill a deer. Yet we will assume that the synergistic effect is smaller than the gains made by defection (i.e. letting someone help you without doing anything in return).

This is realistic if we take into account the fact that the synergy usually only gets its full power after a long term process of mutual cooperation (hunting a deer is a quite time-consuming and complicated business). The prisoner's dilemma is meant to study short term decision-making where the actors do not have any specific expectations about future interactions or collaborations (as is the case in the original situation of the jailed criminals). This is the normal situation during blind-variation-and-selective-retention evolution. Long term cooperations can only evolve after short term ones have been selected: evolution is cumulative, adding small improvements upon small improvements, but without blindly making major jumps.

The problem with the prisoner's dilemma is that if both decision-makers were purely rational, they would never cooperate. Indeed, rational decision-making means that you make the decision which is best for you whatever the other actor chooses. Suppose the other one would defect, then it is rational to defect yourself: you won't gain anything, but if you do not defect you will be stuck with a -10 loss. Suppose the other one would cooperate, then you will gain anyway, but you will gain more if you do not cooperate, so here too the rational choice is to defect. The problem is that if both actors are rational, both will decide to defect, and none of them will gain anything. However, if both would "irrationally" decide to cooperate, both would gain 5 points. This seeming paradox can be formulated more explicitly through the principle of suboptimization.

See also:

- an interactive implementation of the Prisoner's dilemma where you can play the game yourself
- Bjoern Brembs' review on the iterated Prisoner's Dilemma:
- Heylighen F. (1992) : "Evolution, Selfishness and Cooperation", Journal of Ideas, Vol 2, # 4, pp 70-76.