The Prisoner's Dilemma

The Prisoner's Dilemma is a concept in game theory that illustrates a situation where individuals, acting in their self-interest, can lead to a suboptimal outcome for both parties. It is often used to analyze and understand decision-making in various fields, including economics, political science, and psychology.

The classic scenario involves two suspects, A and B, who are arrested and charged with a crime. The police lack sufficient evidence to convict them on the main charge but have enough to convict them on a lesser offense. The suspects are held in separate cells and cannot communicate with each other. The prosecutor makes each suspect the following offer:

If both suspects remain silent, they will each serve a short sentence for the lesser offense (let's say 1 year).

If one remains silent and the other confesses (defects), the one who confesses will receive a reduced sentence (let's say 6 months), while the other will receive a harsher sentence (let's say 5 years).

If both confess, they will each receive a moderately harsh sentence (let's say 3 years).

The dilemma arises from the fact that each suspect must decide whether to cooperate (remain silent) or defect (confess) without knowing the other's choice. The outcomes can be summarized in a payoff matrix:

markdown

Copy code

 | B Stays Silent | B Confesses

-----------------------------------------------

A Stays Silent | -1, -1 | -5, 0

-----------------------------------------------

A Confesses | 0, -5 | -3, -3

The numbers in the matrix represent the respective sentences (in years) for suspects A and B. The first number in each pair represents A's sentence, and the second number represents B's sentence.

The dilemma is that, from a self-interested perspective, each suspect is motivated to confess (defect) to minimize their own sentence. However, if both follow this logic and confess, the overall outcome is worse for both compared to if they had both remained silent.

The Prisoner's Dilemma serves as a model to analyze situations where individuals face conflicting incentives between cooperation and self-interest, leading to outcomes that are collectively suboptimal. It has applications in various fields to understand behavior in strategic interactions.