What Is the Prisoner's Dilemma?

The prisoner's dilemma is a paradox in decision analysis in which two individuals acting in their own self-interests do not produce the optimal outcome.

A prime example of game theory, the prisoner's dilemma was developed in 1950 by RAND Corporation mathematicians Merrill Flood and Melvin Dresher during the Cold War (but later given its name by the game theorist Alvin Tucker).

Today, the prisoner's dilemma is a paradigmatic example of how strategic thinking between individuals can lead to suboptimal outcomes for both players.

- A prisoner's dilemma is a situation where individual decision-makers always have an incentive to choose in a way that creates a less than optimal outcome for the individuals as a group.
- The prisoner's dilemmas occur in many aspects of the economy.
- In the classic prisoner's dilemma, individuals receive the greatest payoffs if they betray the group rather than cooperate.
- If games are repeated, it is possible for each player to devise a strategy that rewards cooperation.
- People have developed many methods of overcoming prisoner's dilemmas to choose better collective results despite apparently unfavourable individual incentives.

The typical prisoner's dilemma is set up in such a way that both parties choose to protect themselves at the expense of the other participant. As a result, both participants find themselves in a worse state than if they had cooperated with each other in the decision-making process. The prisoner's dilemma is one of the most well-known concepts in modern game theory.

The prisoner’s dilemma presents a situation where two parties, separated and unable to communicate, must each choose between cooperating with the other or not. The highest reward for each party occurs when both parties choose to co-operate.

The classic prisoner’s dilemma goes like this:

- Two bank robbers, Elizabeth and Henry, have been arrested and are being interrogated in separate rooms.
- The authorities have no other witnesses, and can only prove the case against them if they can convince at least one of the robbers to betray their accomplice and testify to the crime.
- Each bank robber is faced with the choice to cooperate with their accomplice and remain silent or to defect from the gang and testify for the prosecution.
- If they both co-operate and remain silent, then the authorities will only be able to convict them on a lesser charge resulting in one year in jail for each (1 year for Elizabeth + 1 year for Henry = 2 years total jail time).

- If one testifies and the other does not, then the one who testifies will go free and the other will get five years (0 years for the one who defects + 5 for the one convicted = 5 years total).
- However, if both testify against the other, each will get three years in jail for being partly responsible for the robbery (3 years for Elizabeth + 3 years for Henry = 6 years total jail time).

The respective penalties can be expressed visually as follows:

Possible Outcomes of Prisoner's Dilemma		
Outcome	**Henry Cooperates**	**Henry Defects**
Elizabeth Cooperates	(1,1)	(5,0)
Elizabeth Defects	(0,5)	(3,3)

In this case, each robber always has an incentive to defect, regardless of the choice the other makes. From Elizabeth's point of view, if Henry remains silent, then Elizabeth can either co-operate with Henry and do a year in jail, or defect and go free. Obviously, she would be better off betraying Henry in this case. On the other hand, if Henry defects and testifies against Elizabeth, then her choice becomes either to remain silent and do five years or to talk and do three years in jail. Again, obviously, she would prefer to do the three years over five.

In both cases, whether Henry cooperates with Elizabeth or defects to the prosecution, Elizabeth will be better off if she defects and testifies. Now, since Henry faces the exact same set of choices, he also will always be better off defecting as well.

The paradox of the prisoner's dilemma is this: both robbers can minimize the total jail time that the two of them will do only if they both cooperate and stay silent (two years total), but the incentives that they each face separately will always drive them each to defect and end up doing the maximum total jail time between the two of them of six years total.

Escape from the Prisoner's Dilemma

- A true prisoner's dilemma is typically played only once or else it is classified as an iterated prisoner's dilemma. In an iterated prisoner's dilemma, the players can choose strategies that reward cooperation or punish defection over time. By repeatedly interacting with the same individuals, we can even deliberately move from a one-time prisoner's dilemma to a repeated prisoner's dilemma.
- Collective action to enforce cooperative behavior through reputation, rules, laws, democratic or other collective decision-making procedures, and explicit social punishment for defections transforms many prisoner's dilemmas toward the more collectively beneficial cooperative outcomes.
- some people and groups of people have developed psychological and behavioral biases over time such as higher trust in one another, long-term future orientation in repeated interactions, and inclinations toward positive reciprocity of cooperative behavior or negative reciprocity of defecting behaviors

What Is the Likely Outcome of a Prisoner's Dilemma?

The likely outcome for a prisoner's dilemma is that both players defect (i.e., behave selfishly), leading to suboptimal outcomes for both. This is also the Nash Equilibrium, a decision-making theorem within game theory that states a player can achieve the desired outcome by not deviating from their initial strategy. The Nash equilibrium in this example is for both players to betray one other, even though mutual cooperation leads to a better outcome for both players; however, if one prisoner chooses mutual cooperation and the other does not, one prisoner's outcome is worse.