The Utilitarian and Altruist Prisoner's Dilemma

Natnael M. Yohannes

10 October 2026

The prisoner's dilemma is perhaps the most famous thought experiment in game theory, created by Flood and Dresher around 1950, and later named by Albert Tucker. The game explores an interaction between rational selfish players; however, changing the players' goals can drastically change the conclusions.

Informally, the game consists of two players, each of which may choose to either defect or cooperate. If both players cooperate, they both see mediocre gains. If both defect, they both see minimal gains. If one defects and one cooperates, the defector sees incredible gains and the cooperator receives nothing. The "prisoner's" analogue imagines the scenario where two individuals are accused of a crime, and can either confess or stay silent. Two confessions will result in mutually moderate sentences, two silences result in mutually long sentences, and one confession with one silence results in the silent individual going free while the confessed faces a very long sentence.

Formally, the classic prisoner's dilemma which will be examined is a two-player positive-sum game with two possible choices per individual: *D* for defect and *C* for cooperation. The payoff matrix is shown in fig. 1.

$R(X_1, X_2)$	C_2	D_2
C_1	(3, 3)	(0, 5)
D_1	(5, 0)	(1, 1)

fig. 1

$R(X_1, X_2)$ returns a tuple (Y_1, Y_2), where X_n is the choice of player n and Y_n is the reward to player n.

Traditionally, the prisoner's dilemma is played with perfectly rational agents, each with the sole goal of maximizing their own reward. The outcome of this is that, in an isolated game, agents should always opt to defect. Informally, this is because, regardless of the opponent's decision, defecting always yields a higher reward. Formally, the expected value of each choice is calculated as shown:

$$E_C = (R(C, C).Y_1)P_C + (R(C, D).Y_1)P_D$$
$$E_C = 3P_C$$
$$E_D = (R(D, C).Y_1)P_C + (R(D, D).Y_1)P_D$$
$$E_D = 5P_C + P_D$$

where E_X represents the expected value of an agent making choice X, and P_X represents the probability that the opponent will make choice X. Since the opponent must make either choice, we know that:

$$P_C + P_D = 1$$

Because P_X is a probability between zero and one inclusive, it can be shown that:

$$E_D - E_C = 2P_C + P_D > 0$$
$$\therefore E_D > E_C$$

for any P_C and any P_D. This means that to maximize reward, choosing to defect is always the best option. Even knowing that rational players will always defect by this line of reasoning, setting P_D to one and P_C to zero, it is still more advantageous to defect. In other words, knowing what is optimal does not change what is optimal. We have reached a Nash equilibrium.

A premise of the original game is that players are selfish. By changing the goals of the players, differences in optimal strategy occur. The two examples which will be examined are altruistic and utilitarian players.

Let an altruistic rational agent have the goal of maximizing the opponent's reward, and a utilitarian rational agent the goal of maximizing total reward across both players. These are of course idealized and oversimplified definitions, but serve as examples rather than the exhaustive account of all non-selfish players. The expected values for the altruistic agent are as shown:

$$E_C = (R(C, C).Y_2)P_C + (R(C, D).Y_2)P_D$$

$$E_C = 3P_C + 5P_D$$

$$E_D = (R(D, C).Y_2)P_C + (R(D, D).Y_2)P_D$$

$$E_D = P_D$$

Knowing that

$$P_C + P_D = 1$$

it can be shown that:

$$E_C - E_D = 3P_C + 4P_D > 0$$

$$\therefore E_C > E_D$$

for any P_C and any P_D. This means that the altruistic player is always better off cooperating, regardless of if the opponent defects.

Let a utilitarian rational agent have the goal of maximizing the sum of the opponent's reward and their own reward. The expected values for utilitarian agents are as shown:

$$E_C = (R(C, C).Y_1 + R(C, C).Y_2)P_C + (R(C, D).Y_1 + R(C, D).Y_2)P_D$$

$$E_C = 6P_C + 5P_D$$

$$E_D = (R(D, C).Y_1 + R(D, C).Y_2)P_C + (R(D, D).Y_1 + R(D, D).Y_2)P_D$$

$$E_D = 5P_C + 2P_D$$

it can be shown that:

$$E_C - E_D = P_C + 3P_D > 0$$

$$\therefore E_C > E_D$$

for any P_C and any P_D.

The original conclusion is completely flipped when giving rational agents different goals.

Though the game and payoff matrix stayed the same, optimal strategy still changed as a function of an agent's goal. This example demonstrates narrowly how game-theoretic conclusions differ significantly when considering players beyond traditional selfish agents.

References

Flood, M. M. (1958). Some experimental games. *Management Science, 5* (1), 5–26. https://doi.org/10.1287/mnsc.5.1.5

Nash, J. F. (1950). Equilibrium points in n-person games. *Proceedings of the National Academy of Sciences, 36* (1), 48–49. https://doi.org/10.1073/pnas.36.1.48

Tucker, A. W. (1980). On jargon: The prisoner's dilemma. *The UMAP Journal, 1*, 101.

© 2025 Natnael M. Yohannes