Ben Dupré    50 Philosophy Ideas

# The prisoner's dilemma

***'That's the deal: own up yourself and testify against your mate – he'll go down for 10 years and you just walk away.' Gordon knew the police could send them down for one year anyway, just for carrying the knives; but they didn't have enough to pin the robbery on them. The catch was that he also knew they were cutting the same deal with Tony in the next cell – if they both confessed and incriminated each other, they would each get five years. If only he knew what Tony was going to do . . .***

*. . . Gordon is no fool, so he carefully weighs up his options. 'Suppose Tony keeps quiet; then my best move is to inform on him – he'll get 10 years and I'll go free. And suppose he points the finger at me: it's still best to confess, inform against him, and get five years – otherwise, if I keep quiet, it'll be me doing the 10-year stretch. So either way, whatever Tony does, my best move is to confess.' The problem for Gordon is that Tony is no fool either and reaches exactly the same conclusion. So they incriminate each other and both get five years. Yet if neither had said anything, they would only have got one year each . . .*

So the two men make a rational decision, based on a calculation of their own interest, and yet the outcome is clearly not the best available for either of them. What went wrong?

**Game theory** The story outlined above, known as the 'prisoner's dilemma', is probably the most celebrated of a number of scenarios studied in the field of game theory. The object of game theory is to analyse situations of this kind, where there is a clear conflict of interests, and to determine what might count as a rational strategy. Such a strategy, in this context, is one that aims to maximize one's own advantage and will involve either working with an opponent ('cooperation', in game-theory terms) or betraying him ('defection'). The assumption is, of course, that such analysis casts light on actual human behaviour – either explaining why people act as they do or prescribing how they ought to act.

## Zero sum

Game theory has proved such a fertile field that some of its terminology has become common currency. A 'zero-sum game', for instance – often used informally, especially in business contexts – is technically a game such as chess or poker, where the winnings on one side are exactly balanced by the losses on the other, so the sum of the two is zero. In contrast, the prisoner's dilemma is a 'non-zero-sum' game, where it is possible for both players to win – and for both to lose.

In a game-theory analysis, the possible strategies open to Gordon and Tony can be presented in a 'payoff matrix', as follows:

| | **Tony stays silent** | **Tony confesses** |
|---|---|---|
| **Gordon stays silent** | Both serve 1 year<br>(win–win) | Gordon serves 10 years<br>Tony goes free<br>(lose big–win big) |
| **Gordon confesses** | Gordon goes free<br>Tony serves 10 years<br>(win big–lose big) | Both serve 5 years<br>(lose–lose) |

The dilemma arises because each prisoner is only concerned about minimizing his own jail term. In order to achieve the best outcome for both individuals collectively (each serving one year), they should collaborate and agree to forego the best outcome for each of them individually (going free). In the classic prisoner's dilemma, such collaboration is not allowed, and in any case they would have no reason to trust each other not to renege on the agreement. So they adopt a strategy that precludes the best outcome collectively in order to avoid the worst outcome individually, and end up with a non-optimal outcome somewhere in the middle.

**Real-world implications** The broad implications of the prisoner's dilemma are that selfish pursuit of one's own interest, even if rational in some sense, may not lead to the best outcome for oneself or others; and hence that collaboration (in certain circumstances, at least) is the best policy overall. How do we see the prisoner's dilemma playing out in the real world?

The prisoner's dilemma has been especially influential in the social sciences, notably in economics and politics. It may, for instance, give insight into the decision-making and psychology that underlie escalations in arms procurement between rival nations. In such situations, it is clearly beneficial in principle for the parties concerned to reach agreement on limiting the level of arms expenditure, but in practice they rarely do. According to the games-theory analysis, the failure to reach an agreement is due to fear of a big loss (military defeat) outweighing a relatively small win (lower military expenditure); the actual outcome – neither the best nor the worst available – is an arms race.

A very clear parallel with the prisoner's dilemma is seen in the system of plea bargaining that underpins some judicial systems (such as in the US) but is forbidden in others. The logic of the prisoner's dilemma suggests that the rational strategy of 'minimizing the maximum loss' – that is, agreeing to accept a lesser sentence or penalty for fear of receiving a greater one – may induce innocent parties to confess and testify against each other. In the worst case, it may lead to the guilty party readily confessing their guilt while the innocent one continues to plead their innocence, with the bizarre consequence that the innocent party receives the more severe penalty.

### A beautiful mind

The most famous game theorist today is Princeton's John Forbes Nash. His mathematical genius and triumph over mental illness, culminating in a Nobel Prize for economics in 1994, are the subject of the 2001 film *A Beautiful Mind.*

As a game theorist, Nash's best-known contribution is defining the eponymous 'Nash equilibrium' – a stable situation in a game in which no player has any incentive to change their strategy unless another player changes theirs. In the prisoner's dilemma, double defection (both players confess) represents the Nash equilibrium which, as we have seen, does not necessarily correspond to the optimal outcome for the players involved.

**Chicken** Another game much studied by game theorists is 'chicken', which featured most memorably in the 1955 James Dean film *Rebel Without a Cause*. In the the game, two players drive cars towards each other and the loser (or chicken) is the one who swerves out of the way. In this scenario, the price of cooperation (swerving and losing face) is so small relative to the price of defection (driving straight and crashing) that the rational move appears to be to cooperate. The danger comes when player A assumes that player B is similarly rational and will therefore swerve, thus allowing him (player A) to drive straight with impunity and win.

The danger inherent in chicken is obvious – double defection (both drive straight) means a certain crash. The parallels with various kinds of real-world brinksmanship (potentially most calamitous, nuclear brinksmanship) are equally clear.

**the condensed idea**
**Playing the game**