Introduction

Martin Peterson

0.1 An ingenuous example

The Prisoner's Dilemma is one of the most fiercely debated thought experiments in philosophy and the social sciences. Unlike many other intellectual puzzles discussed by academics, the Prisoner's Dilemma is also a type of situation that many of us *actually encounter* in real life from time to time. Events as diverse as traffic jams, political power struggles, and global warming can be analyzed as Prisoner's Dilemmas.

Albert W. Tucker coined the term "Prisoner's Dilemma" during a lecture in 1950 in which he discussed the work of his graduate student John F. Nash.[1] Notably, Nash won the Nobel Prize in Economics in 1994 and is the subject of the Hollywood film *A Beautiful Mind* (which won four Academy Awards). If this is the first time you have come across the Prisoner's Dilemma, I ask you to keep in mind that the following somewhat artificial example is just meant to illustrate a much more general phenomenon:

> Two gangsters, Row and Col, have been arrested for a serious crime. The district attorney gives them one hour to either confess or deny the charges. The district attorney, who took a course in game theory at university, explains that if both prisoners confess, each will be sentenced to ten years in prison. However, if one confesses and the other denies the charges, then the prisoner who confesses will be rewarded and get away with serving just one year. The other prisoner will get twenty years. Finally, if both prisoners deny the charges, each will be sentenced to two years. The prisoners are kept in separate rooms and are not allowed

[1] Nash writes: "It was actually my thesis adviser, who dealt with my thesis in preparation (on 'Non-Cooperative Games') who introduced the popular name and concept by speaking at a seminar at Stanford U. (I think it was there) while he was on an academic leave from Princeton. And of course this linked with the fact that he was naturally exposed to the ideas in 'Non-Cooperative Games' (my thesis, in its published form)." (Email to the author, December 14, 2012.)

		COL	
		Deny	Confess
ROW	Deny	–2, –2	–20, –1
	Confess	–1, –20	–10, –10

Figure 0.1 The Prisoner's Dilemma

to communicate with each other. (The significance of these assumptions will be discussed at the end of this section.)

The numbers in Figure 0.1 represent each prisoner's evaluation of the four possible outcomes. The numbers −1, −20 mean one year in prison for Row and twenty years for Col, and so on. Naturally, both prisoners prefer to spend as little time in prison as possible.

The Prisoner's Dilemma has attracted so much attention in the academic literature because it seems to capture something important about a broad range of phenomena.[2] Tucker's story is just a colorful illustration of a general point. In order to understand this general point, note that both Row and Col are rationally required to confess their crimes, *no matter what the other player decides to do*. Here is why: If Col confesses, then ten years in prison for Row is better than twenty; and if Col denies the charges, then one year in prison is better for Row than two. By reasoning in analogous ways we see that Col is also better off confessing, regardless of what Row decides to do. This is somewhat counterintuitive, because both prisoners know it would be better for both of them to deny the charges. If Row and Col were to deny the charges, they would each get just two years, which is better than ten. The problem is that as long as both prisoners are fully rational, there seems to be no way for them to reach this intuitively plausible conclusion.

The general lesson is that whenever two or more players interact and their preferences have a very common and reasonable structure, the actions that most benefit each individual *do not benefit the group*. This makes the Prisoner's Dilemma relevant to a broad range of social phenomena. When I do what is best for me, and you do what is best for you, we end up in a situation that is *worse for both of us*. The story of the two prisoners is just a tool for illustrating this point in a precise manner.

[2] Note, however, that some scholars think this attention is unwarranted; see Binmore's contribution to this volume.

We cannot avoid the Dilemma, at least not in a straightforward way, by allowing the prisoners to communicate and coordinate their actions. If Col and Row each promises the other that he will deny the charges, it would still be rational for both men to confess, given that the numbers in Figure 0.1 represent *everything* that is important to them. When the district attorney asks the players to confess, they no longer have a rational reason to keep their promises. If Row confesses and Col does not, then Row will get just one year, which is better than two. It is also better for Row to confess if Col confesses. Therefore, it is better for Row to confess *irrespective of* what Col does. And because the game is symmetric, Col should reason exactly like Row and confess too.

If keeping a promise is considered to be valuable for its own sake, or if a prisoner could be punished for not keeping a promise, then the structure of the game would be different. By definition, such a modified game would no longer qualify as a Prisoner's Dilemma. These alternative games, also studied by game theorists, are less interesting from a theoretical point of view. In this volume, the term "Prisoner's Dilemma" refers to any game that is structurally equivalent to that depicted in Figure 0.1.[3]

For an alternative and perhaps more realistic illustration of the Prisoner's Dilemma, consider two competing car manufacturers: Row Cars and Col Motors. Each company has to decide whether to sell their cars for a high price and make a large profit from each car sold, or lower the price and sell many more vehicles with a lower profit margin. Each company's total profit will depend on whether *the other company* decides to set its prices high or low. If both manufacturers sell their cars at high prices, each will make a profit of $100 million. However, if one company opts for a low price and the other for a high price, then the latter company will sell just enough cars to cover its production costs, meaning that the profit will be $0. In this case, the other company will then sell many more cars and make a profit of $150 million. Finally, if both manufacturers sell their cars at low prices, they will sell an equal number of cars but make a profit of only $20 million. See Figure 0.2.

Imagine that you serve on the board of Row Cars. In a board meeting you point out that *irrespective of what Col Motors decides to do*, it will be better for your company to opt for low prices. This is because if Col Motors sets its price low, then a profit of $20M is better than $0; and if Col Motors sets its price high, then a profit of $150M is better than $100M. Moreover, because

[3] Note that a more precise definition is stated in Section 0.2 of this chapter.

		Col Motors	
		High Price	Low Price
Row Cars	High Price	$100M, $100M	$0M, $150M
	Low Price	$150M, $0M	$20M, $20M

Figure 0.2 Another illustration of the Prisoner's Dilemma

the game is symmetric, Col Motors will reason in the same way and also set a low price. Therefore, both companies will end up making a profit of $20M each, instead of $100M.

The conclusion that the two companies will, if rational, opt for low prices is not something we have reason to regret. Not all Prisoner's Dilemmas are bad for ordinary consumers. However, for Row Cars and Col Motors it is no doubt unfortunate that they are facing a Prisoner's Dilemma. If both companies could have reached a *binding agreement* to go for high prices, both companies would have made much larger profits ($100M). This might explain why government authorities, in protecting consumers' interests, do their best to prevent cartels and other types of binding agreements about pricing.[4]

0.2 Some technical terms explained

Let us try to formulate the Prisoner's Dilemma using a more precise vocabulary. Consider Figure 0.3. By definition, the game depicted in this figure is a Prisoner's Dilemma if outcome A is preferred to B, and B is preferred to C, and C is preferred to D. (That is, $A > B > C > D$.) For technical reasons, we also assume that $B > (A + D) / 2$.[5]

In its classic form, the Prisoner's Dilemma is a *two-player*, *non-cooperative*, *symmetric*, *simultaneous-move* game that has only one *Nash equilibrium*. The italicized terms in the foregoing sentence are technical terms with very precise meanings in game theory.

[4] Some scholars question this type of explanation; see the contribution by Northcott and Alexandrova for a detailed discussion.

[5] This assumption is needed for ensuring that the players cannot benefit more from alternating between cooperative and non-cooperative moves in repeated games, compared to playing mutually cooperative strategies. (Note that we presuppose that the capital letters denote some cardinal utilities. Otherwise the mathematical operations would be meaningless.)

		COL	
		Cooperate	Do not
ROW	Cooperate	B, B	D, A
	Do not	A, D	C, C

Figure 0.3 The generic Prisoner's Dilemma

A *two-player* game is a game with exactly two players. Many Prisoner's Dilemmas are two-player games, but some have three, one hundred, or *n* players. Consider global warming, for instance. I prefer to emit a lot of carbon dioxide irrespective of what others do (because this enables me to maintain my affluent lifestyle), but when all *n* individuals on the planet emit huge amounts of carbon dioxide, because this is the best strategy for each individual, that leads to global warming and other severe problems for all of us.

A *non-cooperative* game is a game in which the players are unable to form binding agreements about what to do. Whether the players actually cooperate or not is irrelevant. Even if the players promise to cooperate with each other, the game would still be a non-cooperative game as long as there is no mechanism in place that forces the players to stick to their agreements. In a non-cooperative game, the players can ignore whatever agreement they have reached without being punished.

That the Prisoner's Dilemma is a *symmetric* game just means that all players are faced with the same set of strategies and outcomes, meaning that the identity of the players is irrelevant. Symmetric games are often easier to study from a mathematical point of view than non-symmetric ones.

That a game is a *simultaneous-move* game means that each player makes her choice without knowing what the other player(s) will do. It is thus not essential that the players make their moves at exactly the same point in time. If you decide today what you will do tomorrow without informing me, the game will still be a simultaneous-move game as long as I also make my move without informing you about it in advance.

The Prisoner's Dilemma is sometimes a simultaneous-move game, but it can also be stated as a *sequential* game in which one player announces his move before the other. Figure 0.4 illustrates a sequential Prisoner's Dilemma in which Player 1 first chooses between two strategies C ("cooperate") and D ("defect"), which is followed by Player 2's choice. The outcome (A1, A2) means that Player 1 gets something worth A1 to him and Player 2 gets A2 units of value. As long as $A1 > B1 > C1 > D1$ and $A2 > B2 > C2 > D2$

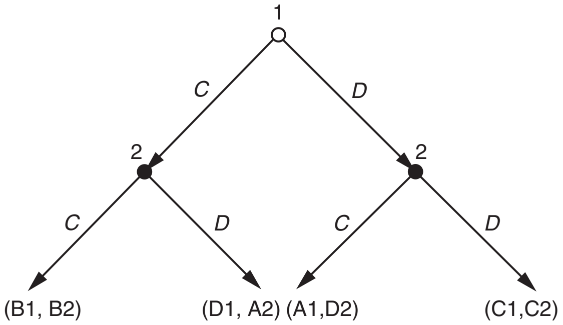

Figure 0.4 In a sequential Prisoner's Dilemma it holds that A1 > B1 > C1 > D1 and A2 > B2 > C2 > D2

the dominance reasoning that drives single-shot versions of the Prisoner's Dilemma will go through.

As explained above, both players' non-cooperative strategies dominate their cooperative strategies in a Prisoner's Dilemma. This type of game therefore has only one *Nash equilibrium*. A Nash equilibrium is a situation in which no player has any reason to unilaterally switch to any other strategy. In his doctoral dissertation, Nash defined the equilibrium concept that bears his name in the following way:

> An equilibrium point is [a set of strategies] such that each player's . . . strategy maximizes his pay-off if the strategies of the others are held fixed. Thus each player's strategy is optimal against those of the others. (Nash 1950: 3)

The key insight is that a pair of strategies is in equilibrium just in case it holds true that *if* these strategies are chosen by the players, then none of the players could reach a better outcome by *unilaterally* switching to another strategy. In other words, rational players will do whatever they can to ensure that they do not feel *unnecessarily* unhappy about their decision, meaning that if a player could have reached a better outcome by *unilaterally* switching to another strategy, the player would have done so.

In order to understand why Nash equilibria are important in game theory, imagine that you somehow knew that your opponent was going to play the cooperative strategy. Would you then, if you were to also play the cooperative strategy, have a reason to unilaterally switch to the non-cooperative strategy ("defect")? The answer is yes, because you would actually gain more by doing so. This shows that if we take Nash's equilibrium concept to be a necessary condition for a plausible principle for how to tackle the Prisoner's Dilemma, then we cannot expect rational players to cooperate.

0.3 The repeated Prisoner's Dilemma

The analysis of the Prisoner's Dilemma depends crucially on how many times it is played. In the single-shot Prisoner's Dilemma a rational player will play the non-cooperative strategy ("defect", i.e. confess the crime), because this strategy dominates the cooperative strategy. Binmore claims in his contribution to this volume that this "is trivial and entirely unworthy of the attention that has been devoted to it."[6] Although Binmore's claim will perhaps not be accepted by everyone, it is important to realize that he is right that it does indeed make a huge difference whether the Prisoner's Dilemma is played only once or many times.

Let us first consider what happens if the Prisoner's Dilemma is repeated (or "iterated") a *finite* number of times. Suppose, for instance, that the Prisoner's Dilemma is repeated exactly three times and that the players know this. To make the example a bit realistic, imagine that each player is a car manufacturer who has exactly three opportunities to adjust the price of some car model during its lifespan. In each round, the companies can choose between a high and a low price. See Figure 0.2. The two players can then reason backwards: In the third round they will know that they are playing the last round; they therefore have no reason to cooperate in that round, meaning that the third and last round can be analyzed as a single-shot Prisoner's Dilemma; each player's non-cooperative strategy dominates her cooperative strategy, so each player will refrain from cooperating in the last round and set a low price.

When the players get to the penultimate round, both players know that neither of them will cooperate in the last round. They therefore have no incentive to cooperate in the penultimate round, because they have no reason to believe that their current behavior will increase the likelihood that the other player will cooperate in the future. The penultimate round can hence be analyzed in the same way as the last one: both players will set their prices low. Finally, the first round can be analyzed exactly like the penultimate round. The players have no reason to cooperate in this round because there is no reason to think their current actions will increase the likelihood that the other players will cooperate in the future.

The argument summarized above is known as the *backward induction argument*. The basic idea is that rational players should reason backwards, from the last round to the first. Note that from a logical point of view it makes no difference if we apply this argument to a Prisoner's Dilemma that is

[6] See Binmore's contribution to this volume, p. 17.

repeated two, three, or a thousand times. However, it is worth keeping in mind that the larger the number of rounds is, the more unintuitive and descriptively implausible the argument will become. Experimental results show that if people get to play a large (but finite) number of rounds, they are likely to cooperate with their opponent because they think this will be rewarded by the opponent in future rounds of the game.[7] This is also how we would expect real-world car manufactures to reason. When they play against each other, they have no pre-defined last round in mind from which they reason backwards. On the contrary, car manufacturers and other large corporations typically seem to think the game they are playing is likely to be repeated in the future, which makes it rational to take into account how one's opponent might respond in the next round to the strategy one is playing now. This indicates that the backward induction argument is a poor analysis of many repeated Prisoner's Dilemmas in the real world.

Many scholars consider the *indefinitely* repeated Prisoner's Dilemma to be the most interesting version of the game. The indefinitely repeated Prisoner's Dilemma need not be repeated infinitely many times. What makes the game indefinitely repeated is the fact that there is no point at which the players *know in advance* that the next round will be the last. The key difference between finitely repeated and indefinitely repeated versions of the Prisoner's Dilemma is thus not how many times the game is actually played, but rather what the players know about the future rounds of the game. Every time the indefinitely repeated game is played, the players know that there is some *non-zero probability that the game will be played again* against the same opponent. However, there is no pre-defined and publicly known last round of the game. Therefore, the backward induction argument cannot get off the ground, simply because there is no point in time at which the players know that the next round will be the last.

So how should rational players behave in the indefinitely repeated Prisoner's Dilemma? The key insight is that each player has reason to take the future behavior of the opponent into account. In particular, there is a risk that your opponent will punish you in the future if you do not cooperate in the current round. In "the shadow of the future," you therefore have reason to cooperate. Imagine, for instance, that your opponent has played cooperative moves in the past. It then seems reasonable to conclude that your opponent is likely to cooperate next time too. To keep things simple, we assume that your opponent has cooperated in the past *because* you have

[7] See Chapter 13.

played cooperative moves in the past. Then it seems foolish to jeopardize this mutually beneficial cooperation by playing the dominant strategy in the current round. If you do so, your opponent will probably not cooperate in the next round. It is therefore better for you to cooperate, all things considered, despite the fact that you would actually be better off in this round by not cooperating.

In the past thirty years or so, game theorists have devoted much attention to indefinitely repeated Prisoner's Dilemmas. The most famous strategy for these games is called *Tit-for-Tat*. Players who play Tit-for-Tat always cooperate in the first round, and thereafter adjust their behavior to whatever the opponent did in the previous round. Computer simulations, as well as theoretical results, show that Tit-for-Tat does at least as well or better than nearly all alternative strategies. Several contributions to this volume discuss the indefinitely repeated version of the Prisoner's Dilemma.

0.4 Overview

This volume comprises fourteen new essays on the Prisoner's Dilemma. The first three chapters set the stage. The next ten chapters zoom in on a number of specific aspects of the Dilemma. The final chapter draws conclusions.

In Chapter 1, Ken Binmore defends two claims. First, he argues that all arguments for cooperating in the single-shot Prisoner's Dilemma proposed in the literature so far are fallacious. Such arguments either alter the structure of the game, or introduce additional, questionable assumptions that we have no reason to accept. The only rational strategy in the non-cooperative version of the game is, therefore, the single-shot strategy. Binmore's second claim concerns the connection between the Prisoner's Dilemma and Kant's categorical imperative. Binmore's conclusion is that although the indefinitely repeated Prisoner's Dilemma can shed light on the evolution of a number of social norms, "the Prisoner's Dilemma shows that [the categorical imperative] can't be defended purely by appeals to rationality as Kant claims." So, according to Binmore, Kant was wrong, at least if the notion of rationality he had in mind was the same as that researched by game theorists.

David Gauthier, in Chapter 2, disagrees with Binmore on several issues. The most important disagreement concerns the analysis of single-shot Prisoner's Dilemmas. Unlike Binmore, Gauthier claims that it is sometimes (but not always) rational to cooperate in a single-shot Prisoner's Dilemma. The argument for this unorthodox view goes back to his earlier work. As Gauthier now puts it, the main premise is that, "if cooperation is, and is recognized by (most) other persons to be possible and desirable, then it is rational for each to be a cooperator." Why? Because each person's own objectives can be more easily realized if they agree to bring about a mutually beneficial outcome. It is therefore rational for each person to think of herself as a cooperator and deliberate together with the other player about what to do.

It should be stressed that Gauthier's argument is based on a theory of practical rationality that differs in important respects from the one usually employed by game theorists. According to the traditional theory, an agent is rational if and only if she can be described as an expected utility maximizer. This, in turn, means that her preferences obey a set of structural conditions (called "completeness," "transitivity," "independence," and "continuity"). It is moreover assumed that what agents prefer is "revealed" in choices, meaning that the agent always prefers what she chooses. Gauthier rejects this traditional picture and sketches an alternative. In Gauthier's view, preference and choice are separate entities, meaning that an agent can make choices that conflict with her preferences. This alternative account of practical rationality, which will probably appeal to many contemporary philosophers, is essential for getting Gauthier's argument off the ground. In order to coherently defend the claim that it is sometimes rational to cooperate in the single-shot Prisoner's Dilemma, the preferences that constitute the game cannot be identified with the corresponding choices.

In Chapter 3, Daniel M. Hausman discusses the notion of preference in the Prisoner's Dilemma. Economists take the player's preference to incorporate all the factors other than beliefs that determine choices. A consequence of this close relation between preference and choice is that it becomes difficult to separate the influence of moral reasons and social norms from other factors that may influence one's choice. If you choose to cooperate with your opponent in what appears to be a Prisoner's Dilemma because you feel you are under a moral obligation to do so, or because your behavior is influenced by some social norm, then you prefer to cooperate, and the game you are playing is not a Prisoner's Dilemma. If, as Hausman believes, cooperation in what appear to be Prisoner's Dilemmas often shows that people are not in fact playing Prisoner's Dilemmas, then the experimental results pose no challenge to the simple orthodox analysis of the Prisoner's Dilemma. But this leaves game theorists with the task of determining what game individuals in various strategic situations are actually playing. Hausman discusses some sophisticated ways of thinking about how to map "game forms" into games, especially in the work of Amartya Sen and Cristina Bicchieri, but he concludes that there is a great deal more to be done.