Philippe van Basshuysen

Book review: the prisoner’s dilemma, Martin Peterson (ed.). Cambridge University Press, 2015, viii + 298 pages.

Article (Accepted version) (Unrefereed)

Original citation:
van Basshuysen, Philippe (2017) *Book review: the prisoner’s dilemma, Martin Peterson (ed.). Cambridge University Press, 2015, viii + 298 pages.* Economics & Philosophy, 33 (1). pp. 153-160. ISSN 0266-2671

DOI: http://dx.doi.org/10.1017/S0266267116000249

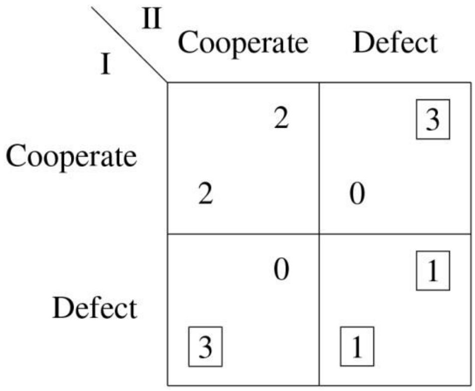
The Prisoner's Dilemma, Martin Peterson (ed.). Cambridge University Press, 2015, viii + 298 pages.

The Prisoner's Dilemma is a collection of discussions by philosophers and economists on the most-debated example from game theory. The Prisoner's Dilemma (henceforth 'PD') is also the prime example for game theory's ability to fuel debates about rationality that have captivated many researchers interested in how agents think, or should think, in interactive situations.

For an illustration of the PD structure, an example of a (one-shot, two player) PD game is depicted in figure [1], in which numbers represent utilities. Players I and II choose between two strategies, "Cooperate" and "Defect". For both players defection yields a higher payoff than cooperation no matter what the opponent does. It follows that no player can gain by unilaterally changing her strategy only if both defect: (Defect, Defect) is the only Nash equilibrium of the PD. However, in equilibrium, the payoffs for both players are smaller (in the example, 1 for both) than what they would have received had both played the dominated, cooperative strategy (2).

What makes the game a "dilemma" is this discrepancy between what may be interpreted as individually rational acts and the socially undesirable outcome they constitute. The PD structure raises questions about the concept of rationality, cooperation, the interpretation of game theory, and its relation to real-world interactions and behaviour.

In section 1 of this review, I give a brief summary of the book and propose a way of thinking about its themes. In section 2, I focus on the dispute about whether it can be rational to cooperate in a one-shot PD. Ken Binmore defends the orthodox view that it is irrational to do so in chapter 1. By contrast, in chapter 2, David Gauthier argues it can be rational. I shall argue that Gauthier's view is internally inconsistent. Thus, if one reads Gauthier's chapter as a challenge to Binmore, it is, if I am right, a failed challenge.



II I	Cooperate	Defect
Cooperate	2 2	$\boxed{3}$ 0
Defect	0 $\boxed{3}$	$\boxed{1}$ $\boxed{1}$

Figure 1: example of a PD game. Player I's payoff is shown on the bottom left and player II's payoff on the top right of each cell. The best responses for the players are marked with squares around the payoff numbers. (Defect, Defect) is an equilibrium in pure strategies (in fact, it is the only equilibrium).

1. Overview and thematic organization.

The book's structure is somewhat haphazard, and one need not read its fourteen chapters in the order in which they are presented. To help the reader, I propose organising them into the following five themes.

(1.) Rational play in the PD. The first theme is whether cooperation can be rational in one-shot PDs. As we have seen, cooperating in such games means playing a dominated strategy instead of playing the equilibrium strategy prescribed by classical game theory – an irrational move, according to the orthodox analysis. Binmore defends the orthodox view. In contrast, Gauthier, Martin Peterson, and Giacomo Bonnano all agree – albeit for different reasons – that it can (although it may not always) be rational to cooperate in such a game.

Gauthier takes Pareto efficiency instead of utility maximisation as the primary goal of a theory of practical rationality. For him, it is a *reductio* of the orthodox analysis that defection which leaves every player worse off is the prescribed strategy of non-cooperative game theory. Even worse, in his view, is the orthodox theory's prescription, using backwards induction, to always defect in finitely repeated PDs with fixed, commonly known endpoints.

Bonnano analyses counterfactual reasoning in one-shot PDs. According to him, a player's beliefs that her opponents will cooperate, conditional on the cooperation of the player herself, can serve as an argument to rationalise the cooperation strategy.

Peterson argues we should model a player as an aggregate of subagents with different aims that play "internal" games against each other, and that "external" games a player plays against other players can be described as being part of internal games. Since subagents will likely encounter each other again in future games of similar structure, this allows modelling external, one-shot PDs as internal, indefinitely repeated PDs. According to Peterson, this may serve as an argument for cooperation in one-shot PDs.

(2.) Conditions for cooperation. The second theme looks at the conditions under which cooperation can evolve, or become a rational strategy (assuming that it is not a rational strategy in the first place – in this respect, the theorists in this group seem to agree with Binmore and disagree with the authors discussed above).

Paul Weirich argues that individual rationality of the members of a group constitutes the group's collective rationality. What the PD shows is that collective rationality does not imply efficiency or maximising collective utility: suboptimal outcomes in PD-like situations are deficiencies but are 'excused' (p. 270) by the PD structure.

The chapters of Jeffrey Barrett and of Cristina Bicchieri and Alessandro Sontuoso focus on communication as enabling cooperation. Barrett discusses cooperation in an evolutionary setting with pre-play signalling. Bicchieri and Sontuoso build on Bicchieri's work on social norms: communication enables agents to focus on social norms, which if complied with, are mutually beneficial. These norms may then guide agents' choices of strategy and, in particular, may lead them to cooperate. They also summarise empirical evidence confirming the hypothesised positive correlation between communication and cooperation.

(3.) Incentives in PDs. Geoffrey Brennan and Michael Brooks investigate players' incentives relative to the total number of players, in both PDs and Public Goods games. It is generally claimed that the incentives to defect in one-shot PDs increase with the number of players. Brennan and Brooks agree, other things being equal, but argue that external effects not occurring in the framing of a problem as a PD – such as the number of observers of players' choices – may even work in the opposite direction.

Charles Holt, Cathleen Johnson and David Schmidtz present new empirical results on how agents play in finitely repeated PDs with commonly known endpoints. Cooperation increases with the length of the game, and when the interaction between the players is voluntary, i.e. there is an exit option to playing the game. These findings may serve as policy advice on how to promote cooperation.

(4.) Applications and applicability. Anna Alexandrova and Robert Northcott challenge the explanatory role of the PD and with it that of game theory more generally. They claim that attempts such as Axelrod's (1984) to force the occurrence of historical events (e.g., World War I truces) into a game theoretic straitjacket (an indefinitely repeated PD) fail in comparison to historical explanations. Moreover, they claim that very few real-world phenomena even have a PD structure.

The next two essays implicitly contradict these claims, modelling social phenomena as PDs or game theoretically more generally. Douglas MacLean argues that the global problem which anthropogenic climate change poses is a (one-shot, because future generations cannot reciprocate) PD: a country's reducing emissions ("cooperating") is dominated by sticking to the status quo ("defecting") because it would have economic disadvantages and the effects for the global climate of unilaterally reducing emissions are small. What makes the enforcement of cooperation difficult is that failing to reduce emissions will affect mainly future generations. There is widespread agreement that there should be discount rates to the costs of future damages and future emission reductions. However, the size of the discount rates is a central controversy that has received special attention since the release of the *Stern Review* (Stern 2007). MacLean sides with Stern that the discount rates should be relatively small.

Luc Bovens argues that classical tragedies of the commons, as they are presented in Aristotle, Hume and Mahanarayan, have the structure of a "voting game" rather than a PD. A voting game, in Bovens' terminology, is an *n*-player noncooperative game with payoffs that are made up of a simple, monotonic coalitional form game where joining a coalition (meaning cooperating, or casting a vote) is costly. It follows that only players in minimal winning coalitions have incentives to do so. The equilibria of such a game are "all defect" and mixed equilibria where a minimal coalition forms and all players outside the minimal coalition defect. The instability of the mixed equilibria may then explain the likely outcome in Tragedy-of-the-Commons-like situations: the suboptimal "all defect" equilibrium.

(5.) Framing/modelling. The applications and applicability theme is intimately connected with the problem of how to model specific games or social structures. José Luis Bermúdez asks whether the PD can be modelled as Newcomb's problem, as was argued by David Lewis in his defence of causal decision theory (Lewis 1979). Bermúdez thinks it cannot. He argues that Lewis commits the same fallacy as can be found in the symmetry argument in favour of cooperating in a one-shot PD. The symmetry argument is this: in a symmetric game with common knowledge of rationality among the players, only symmetric choices are possible and hence cooperation is the only rational strategy. The argument fails because it effectively changes the structure of the game – the resulting game is no longer a one-shot PD.

Daniel Hausman discusses the question of how social situations should be modelled. Many situations which may from the outside appear to have the structure of PDs are not conceived as PDs by players. For example, in many PD experiments, the payoff numbers are monetary payoffs, but the players involved in the game may add secondary utilities for cooperative behaviour. They may appear to be playing a dominated strategy when really they aren't. According to Hausman, the problem of mapping specific situations to game forms is a central worry for game theory, and more work should be dedicated to it.

2. Is cooperation rational in a PD?

The dispute between Gauthier and Binmore merits more attention. Gauthier sketches a theory which advances Pareto efficiency as the decisive criterion for a rational solution to a game, as opposed to the equilibrium criterion which orthodox game theory prescribes. He claims that his theory rationalises cooperation in one-shot PDs and finitely repeated PDs with fixed, commonly known endpoints.

It is instructive to contrast Gauthier's with different views on rationality. Gauthier calls theories of rationality that rely on the equilibrium concept *best reply theories* because a combination of strategies in which each player's action is a best reply to the other players' actions constitutes an equilibrium. According to Binmore's best reply theory, all and only equilibrium strategies (i.e. strategies that are played with positive probability in some equilibrium) are rational. Games may have many equilibria. A stronger best reply theory than Binmore's is one according to which playing an equilibrium strategy is necessary but not sufficient for rationality. Harsanyi and Selten (1988) championed this approach with a general theory that selects a uniquely rational equilibrium for every game.

In contrast, according to Gauthier, playing an equilibrium strategy is neither necessary nor sufficient for rational play. Consider again the PD in figure [1]. Gauthier thinks the only equilibrium (Defect, Defect) cannot be the solution a theory of rationality should prescribe. Two cooperators playing the PD manage to interact in a way that makes both better off than two defectors. What is necessary for rational play is Pareto efficiency, whenever the opponent is likely to cooperate too. If he doesn't, it is rational to fall back on best reply reasoning and defect instead of falling prey to his play. In short, Gauthier seeks to rationalise cooperation in a potentially cooperative environment.

I will argue that Gauthier does not succeed in rationalising cooperation. In a nutshell, my argument is the following. Gauthier's theory – although put forward as a general theory of "practical rationality" – seems to be primarily motivated by the aim of justifying cooperation in the PD. It yields rather absurd recipes for rational behaviour when applied to game structures different from the PD, even for (and perhaps particularly for) theorists sympathetic to cooperation in the PD. To the extent that Gauthier's theory fails as a general theory of rationality, it fails as a theory for prescribing rational behaviour in the PD.

The problem is that his theory which he claims is opposed to best reply theories is really married to best reply theories, and that the marriage is not a happy one. To begin with, note that Pareto efficiency is a weak concept and in many games entirely indecisive. For example, in zero sum games, every possible strategy combination is Pareto efficient. To put teeth into his theory, Gauthier bases it on best reply theories: he requires choosing a Pareto efficient equilibrium if one exists; and otherwise an outcome that Pareto dominates all equilibria (p. 41). But this recipe yields results that conflict with his view, as the next example shows.

Consider an example of the Chicken Game, depicted in figure [2]. The game has three equilibria: the pure equilibria (Straight, Swerve) and (Swerve, Straight), and a mixed equilibrium in which both players play swerve with probability 9/10 and straight with probability 1/10. In the pure equilibria the players receive a payoff of (1, -1) and (-1, 1) respectively. In the mixed equilibrium, both players receive payoff -1/10. The pure equilibria are Pareto efficient whereas the mixed equilibrium is dominated by (Swerve, Swerve) which gives both players payoff 0. Hence, applying Gauthier's recipe, only the pure equilibria are candidates for rational play. I argue that this is a self-defeating outcome for two reasons.

First, it is an outcome that rocks the cooperator's boat. Both pure equilibria leave one of the players in negative figures, and seem precisely to violate cooperators' willingness to find a mutually advantageous outcome. Consider, in contrast, the non-equilibrium compromise (Swerve, Swerve) in which both players receive 0 payoff. This mutually advantageous outcome has recently been suggested by theorists defending the rationality of non-equilibrium play (Karpus and Radzvilas 2016). It has a "cooperative spirit" which the pure strategy equilibria lack. As described above, this "cooperative spirit" is the very motivation for Gauthier's theory; yet the theory is unable to choose the outcome that realises it. The theory's aim and its realisation are thus in conflict.

The claim that (Swerve, Swerve) in the Chicken Game has a "cooperative spirit" can be made more precise. Suppose we transform the game in figure [2] into a cooperative game. In such games, players can reach Pareto efficient outcomes which are not in equilibrium through binding agreements. So in the Cooperative Chicken Game, all three Pareto efficient outcomes (Swerve, Swerve), (Swerve, Straight), and (Straight, Swerve) can be reached. Which one will be reached? According to Nash (1953), the selection of the cooperative outcome can be seen as the solution to a bargaining game. If the players have equal bargaining power, the outcome in our game is precisely (Swerve, Swerve).

I \ II	Swerve	Straight
Swerve	0, 0	$\boxed{-1}$, $\boxed{1}$
Straight	$\boxed{1}$, $\boxed{-1}$	-10, -10

Figure 2: example of a Chicken Game. There are three equilibria: two pure equilibria indicated in the figure; and a mixed equilibrium in which both players play swerve with probability 9/10 and straight with probability 1/10.

Second, Gauthier's theory does not only rely on best reply theories; it relies on a best reply theory that selects a *unique* equilibrium. Moreover, whenever there are Pareto efficient equilibria among the equilibria in the game, the best reply theory must select one among them (p. 41). According to Gauthier, this Pareto efficient equilibrium is the unique rational outcome of the game.

Gauthier does not describe what this best reply theory should look like, and in fact this would be an impossible task. Consider the example of the Chicken Game again. Once the Pareto-dominated mixed equilibrium is eliminated, there remain two Pareto efficient equilibria. However, given the symmetries of the game, there are no rational grounds on which to distinguish between the two equilibria. Hence, no theory of rationality could distinguish between them. For this reason, Harsanyi and Selten's (1988) theory – which always selects a unique equilibrium – selects the mixed equilibrium in this game.

Harsanyi and Selten's theory can be criticised on the grounds that it may select Pareto-dominated equilibria. They cannot be criticised for inconsistency. Binmore's theory can be criticised on the grounds that it deems all equilibria – Pareto efficient or not – equally rational. But he, likewise, cannot be criticised for inconsistency. Gauthier demands that a theory of rationality select a unique, Pareto efficient equilibrium whenever there are Pareto efficient equilibria. This demand is impossible to meet, and so refutes his theory. Note that my argument does not preclude the possibility of there being a consistent theory of rationality which implies cooperation in the PD. Gauthier's, however, is not such a theory.

3. Conclusion

I can recommend the present volume to researchers and students interested in the foundations of game theory and its applications. It is also well-suited to be read in a graduate semester course on the PD. It could be read cover-to-cover, or following the thematic organisation I proposed in section 1. As the discussion in section 2 indicates, the volume comprises lively discussions of at times opposing views. One can hope that these original and up-to-date contributions will stimulate further foundational research in game theory, the philosophy of its application, and the philosophy of rationality.

Philippe van Basshuysen*

ACKNOWLEDGEMENTS

Many thanks to Jason Alexander, Luc Bovens, Goreti Faria, Jurgis Karpus, James Nguyen and Alex Voorhoeve for helpful comments on an earlier draft.

REFERENCES

Axelrod, R. 1984. *The Evolution of Cooperation*. New York: Basic Books.

Harsanyi, J.C. and R. Selten. 1988. *A General Theory of Equilibrium Selection in Games*. Cambridge, MA: The MIT Press.

Karpus, J. and M. Radzvilas. 2016. Team Reasoning and a Measure of Mutual Advantage in Games. Preprint, http://philsci-archive.pitt.edu/12364/.

Lewis, D. 1979. Prisoner's Dilemma Is a Newcomb Problem. *Philosophy and Public Affairs* 8, 235-240.

Nash, J.F. 1953. Two-Person Cooperative Games. *Econometrica* 21(1), 128-140.

Stern, N. 2007. *The Economics of Climate Change: The Stern Review*. Cambridge: Cambridge University Press.

BIOGRAPHICAL INFORMATION

Philippe van Basshuysen is a PhD candidate at the London School of Economics and Political Science. He works on decision and game theory and their application to social science and policy-making, with a focus on market and mechanism design.

*London School of Economics, Houghton Street, London WC2A 2AE, UK. Email: P.C.Van-Basshuysen@lse.ac.uk