GAMES, PUBLIC GOODS, AND THE JUSTIFICATION OF THE STATE

JACK LANG
UNIVERSITY COLLEGE LONDON
MPHIL PHILOSOPHY 2011

DECLARATION

I, Jack Lang, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed _____

ABSTRACT

The normative assessment of the state, I argue, may take two main forms. One may focus on legitimacy, which is established through voluntary and morally significant relationships between individuals and the states that claim authority over them. I focus on a second type of assessment; that of justification. A state, and its coercive acts, may be justified (even if that state is not legitimate) when its existence is, for whatever reason, taken to be beneficial to those it coerces.

One popular justification of the state - the so-called 'public goods argument' - invokes the prisoners' dilemma; the coercive state, on such a view, is justified insofar as it allows individuals to achieve the mutually beneficial outcomes that would evade them in its absence. In the second half of the paper, I argue that such a game theoretical argument fails. Two types of account suggest that mutually beneficial interaction (cooperation) is possible even within the prisoners' dilemma framework. One of those accounts, I argue, is rather more successful than the other. Those positions also highlight reasons for doubting that the game theoretical framework can provide a plausible basis for the justificatory task.

TABLE OF CONTENTS

GAMES, PUBLIC GOODS, AND THE JUSTIFICATION OF THE STATE

INTRODUCTION

In recent times, the use of game theory has become increasingly prevalent in political philosophy. The analytical tools it provides are undoubtedly attractive at first glance, offering clear and lucid explanations of certain behavioural phenomena. Certain games, however, have been taken to form a plausible basis of more normative positions. One type of game in particular – the prisoners' dilemma – has been employed in debates surrounding the state, and the nature of our relationship with it. It is this game, and its use within one type of normative discussion, that forms the subject of this thesis.

The normative assessment of the state, I argue, may take two main forms. Firstly, one may focus on legitimacy, which, I posit, is established through voluntary and morally significant relationships between individuals and the states that claim authority over them. My focus in this paper, however, is on a second type of assessment: that of justification. A coercive state may be justified (even if that state is not legitimate) when its existence is, for whatever reason, taken to be beneficial to those whom it coerces.

One popular justification of the state invokes the prisoners' dilemma; the coercive state, on such a view, is justified insofar as it allows individuals to achieve the mutually beneficial outcomes that would evade them in its absence. In the second half of the paper, I argue that such a game theoretical argument fails. Two types of account suggest that mutually beneficial interaction (cooperation) is possible even within the prisoners' dilemma framework. Those positions also highlight reasons for doubting that the framework can provide a plausible basis for the justificatory task.

CHAPTER 1 – THE STATE, LEGITIMACY, AND JUSTIFICATION

The primary questions in political philosophy concern our relationship with the state. The issue is undoubtedly a complex one, involving descriptive, explanatory, and normative tasks. Our common understanding of such a relationship tends to involve a dense web of notions, such as obligation, duty, legitimacy, justification, authority, power and coercion. The occasional lack of conceptual clarity that surrounds these notions can obscure important distinctions among them.

My task in this chapter is to delineate two different ways of normatively evaluating the state; ways which are often run together. The question of state *legitimacy*, I wish to argue, should be thought of as concerning a state's right to rule; its right, established on the basis of some (morally) relevant interaction with certain individuals, to impose obligations, and to employ coercive measures to enforce them. A distinct issue is that of *justification*; a state, whether legitimate or not, may be justified if it proves effective at providing certain goods, or exhibits certain characteristics, such that its existence is taken (in a manner I shall go on to specify) to be a 'good thing.'

I make this distinction for two reasons. Firstly, I do so in order to clarify the location of my project within the field of political philosophy. My aim in this paper is to present, and subsequently criticise, what I (and others) take to be the most plausible version of a justification of the state. If my argument in the current chapter is clear and convincing, it will thereafter be clear why I don't take such proposals to bear on the question of legitimacy, and those concepts related to it. The distinction made in this chapter will allow me to cordon off a set of questions, which, I take it, do not relate directly to the crux of my thesis.

Secondly, I make the distinction in order to motivate interest in the justificatory position under scrutiny. As well as being simply clarificatory, the distinction in question is an important one. As I shall note, it suggests why we should regard justification as significant, despite being divorced from the (seemingly more fundamental) question of legitimacy. I indicate that, in light of a particular challenge posed to the common assumptions about legitimacy – that of philosophical anarchism – justification can be a key tool in the moral assessment of the state.

Characterising the state

Before I embark on the main task of this chapter, it will be salient to provide a brief characterisation of the state. States, quite evidently, come in all shapes and sizes; both in the real world, and in terms of the hypothetical recommendations of philosophers. Primitively, I think, we view the state as a body that exhibits, at the very least, the following features.

The state, through its directives, seeks to provide individuals with reasons to act in certain ways that are stipulated by those directives. Individuals may be given (new) reasons for action in a number of ways. When I am threatened by a knife-wielding thug on the street, I may gain a reason to act on the way that he orders; I hand over my wallet because the considerable disutility I attach to being stabbed makes that course of action the most rational in that scenario. The state may provide individuals with reasons in this way; it exercises *power* to change the conduct of individuals.

It also, however, goes further than this. Within a certain territory, the state claims legitimate *authority*; what we way may characterise, for the instant, as some right to rule. Legitimate authority, we might think, creates reasons for individuals in a different manner; individuals obey the directives of the state out of a moral duty, not simply out of fear, or due to prudential reasons. The claim to legitimate authority, then, is what distinguishes the state from a knife-wielding thug; the former seeks to impose moral obligations in a manner that the latter does not. Such an effort to establish legitimacy need not be successful, as we shall see presently, but I posit that the attempt is necessary for statehood.[1]

As the famous Weberian characterisation pinpoints, the state generally claims a *monopoly* on the use of legitimate force within a given territory; it claims not only the right to rule, but also that it alone has such a right.[2] Coercive action by any body other than the state itself – or those bodies to whom it

[1]We may follow Joseph Raz here in distinguishing between the notions of *de facto* authority and *de jure* authority. The former, on Raz's account, refers to a body that claims to, but does not have, a right to rule. The latter term equates to legitimacy; a *de jure* authority claims, and actually has, a right to rule. See Raz, 'The Problem of Authority: Revisiting the Service Conception,' in *Minnesota Law Review*, Vol.90, No.4 (April 2006), p1005.

[2] Weber, 'Politics as a Vocation,' in *The Vocation Lectures*, trans. Rodney Livingstone (Hackett, 2004), p33.

delegates such force – is, whether by legal means or otherwise, something that the state seeks either to disallow, or to regulate.

Can a body be an authority without wielding power? It seems fairly clear that it can. Authority may be – to employ terms used by Raz – theoretical rather than practical; it may provide individuals with reasons for belief, rather than reasons for action.[3] On this view, authority is linked to expertise; we may take certain individuals or groups of individuals to be authorities on certain issues or within certain fields. Listening to such authorities, heeding their advice with respect to particular topics, provides us with "information about the balance of reasons as they exist separately and independently" of their utterances.[4] We do not, however, require such authorities to wield (or even attempt to wield) power; advice is not something with which we generally expect to be forced to comply. In addition, most theoretical authorities seem unlikely to press a claim for legitimacy; in fact, it seems doubtful whether the very notion of legitimacy applies to this type of authority at all.

This type of case, however, should not guide our general characterisation of the state. The state, I take it, is a paradigm case of practical authority; it is a body which provides individuals with reasons for action, rather than for mere belief. The utterances of the state generally take the form, not of advice, but of commands. The directives themselves, backed up by coercion, are intended to create new reasons for individuals to act (or refrain from acting) in certain ways, rather than merely to flag up existing reasons. That the word 'authority' can stretch to cover expertise, then, should not obscure an important distinction. Richard Flathman characterises the difference as one between "in authority" and "an authority"; the latter notion, which would be used in sentences of the type "Anne is an authority on breeds of dog," is simply not relevant to the study of the state, which focuses on whoever is *in* authority.[5]

Does, though, the fact that states are practical authorities – bodies that have the ability to alter the balance of reasons of putative subjects – require them to wield power? Could a state not provide reasons for action merely through the morally binding nature of its directives? I do not deny this possibility; a

[3] Raz, 'Authority and Consent,' in *Virginia Law Review*, Vol.67: 103 (1981), p108.
[4] Ibid, p108.
[5] Flathman, *The Practice of Political Authority* (University of Chicago Press, 1980), p16-17.

legitimate state, ruling over individuals who unfailingly acted on their obligations, might secure compliance without the need for coercion, or even the threat of coercion. Such a hypothetical situation, though, seems to bear very little on the issues at hand.

Firstly, individuals are not morally perfect; they are tempted to, and indeed do, break their obligations, for all manner of reasons. Given this (admittedly primitive) fact about human psychology, the use of coercion would seem to be generally required to secure compliance. Secondly, for reasons I shall come on to, we may doubt that many (or, indeed, any) legitimate states exist. If this were the case, the likelihood of the kind of moral practical authority outlined above would be little or none. Again, it would seem that the wielding of coercive power is a general (if not strictly necessary) feature of states.

We should also, at this point, bring out a distinction between states and specific political bodies, such as governments. Evidently, the two are closely related; governments and the like are commonly the most visible manifestations and instruments of the state, particularly at the current time. Yet the term 'state,' I think, should be considered a broader one; one that might, but need not, evoke specific types of institutions. The existence of a state, on this view, doesn't simply equate to the existence of a particular type of governance at a given instant; rather, a state may be said to exist wherever, and as long as, *some* structure or framework claims a monopoly on the use of legitimate force within a certain territory.

With this clarification in mind, we might think that the term 'legitimate' applies to states and specific governments differently. State legitimacy, I shall argue, is based on some voluntary, obligation-generating interaction between the state and the individuals over whom it claims authority. Governmental legitimacy seems to arise as a secondary question in legitimate states, one that relates to some procedure or due process. As such, it seems right to say that an illegitimate government can exist within a legitimate state (but not vice versa).

This position seems to be consonant with a (broadly) Lockean framework. A. John Simmons presents a plausible reading of Locke, whereby;

> …states […] earn their legitimacy by virtue of the (unanimous) consent of their members, a consent that transfers to the collectivity those rights

> whose exercise by a central authority is necessary for a viable political society. Governments are legitimate only if they have been entrusted by the state (society) with the exercise of those same rights. So while a legitimate state might have an illegitimate government (one that, say, acquired its power by force rather than by trust), an illegitimate state could never have a legitimate government since illegitimate states do not possess the rights, transferred to them by their subjects' consents, that must be entrusted by a state to a government in order to legitimate that government.[6]

As shall become clear, I am sympathetic to such a Lockean view. Note, however, that the distinction between state and governmental legitimacy needn't invoke convictions about the transfer or entrustment of rights. A legitimate government, on a common sense view, would simply be one that has been instituted according to some constitutional process, such as a free, democratic election. I neither venture an account of governmental legitimacy here, nor appraise the candidate views. Suffice it to reiterate at this juncture that I take the question of state legitimacy to be more fundamental than, and thus distinct from, the issue of governmental legitimacy.

Clearly, there is potential for debate about what *type* of governance is legitimate, preferable, feasible, and so on; these, I take it, are questions about *what form* the state should take. This paper, however, invokes what I take to be a more basic question: why have a state at all? This latter enquiry may seem to be a nebulous one, because it may initially be unclear what the alternative is. Whilst we are acquainted with the idea that various different types of rule exist, and that we may be able to choose the type that we prefer (either by participating in the political process, or by deciding which state to live in), experience does not furnish us with many tools to elaborate the possibility of not living in a state at all. States are now (and have long been) a sociological given; we are born within them, and can escape their grasp only by making a sustained effort, if at all.[7]

There may be myriad responses to the question: "why a state?" One answer forms the focus of this thesis. Before I present it, however, it is necessary

[6] Simmons, 'Justification and Legitimacy,' in *Ethics*, Vol. 109, No. 4 (July 1999), p747.
[7] By, for instance, moving to a remote island that is not (yet) under the control of any state.

to distinguish between two different types of moral evaluation that the state may be submitted to; namely, assessments of legitimacy and of justification. The latter refers primarily to the features the state exhibits, such as its efficacy in providing certain goods. I take the account offered in the next chapter to be an instantiation of this method of evaluation. The former, meanwhile, concerns certain morally significant ways in which the state interacts with the individuals over whom it claims authority. It is to this type of assessment, and the closely related notion of political obligation, that I now turn.

Legitimacy and political obligation

Legitimacy, as I have primitively defined it thus far, is a right to rule; legitimate authorities somehow have some moral dispensation to issue directives and enforce them, in a way that mere de facto authorities do not. There is a huge philosophical literature that centres upon the question of legitimacy, as well as that of political obligation; indeed, the two notions are often treated in tandem, for reasons that I shall elucidate below.

States tend to (or at least seek to) impose obligations – requirements to act in certain ways – upon their putative subjects. Such requirements are intended to override other reasons that may apply to the individual in a given situation; obligations are to be discharged "regardless of our inclinations."[8] Perhaps the most discussed of these obligations is the alleged obligation to obey the law; individuals are required to conduct themselves in a manner compatible with the guidelines laid down by the legal framework of the state, even if such behaviour conflicts with their individual preferences.

H.L.A. Hart's characterisation of obligations is an instructive one. For Hart, an obligation is a moral requirement that satisfies four conditions;

(1) An obligation is created by the performance of a voluntary act (or omission).
(2) An obligation is owed by a specific actor to a specific person/set of persons.

[8] Simmons, *Moral Principles and Political Obligations* (Princeton University Press, 1979), p7.

(3) An obligation is content-independent.

(4) An obligation simultaneously generates a correlative right.[9]

These features, particularly (1) and (2), point to ways in which obligation is distinct from duty. Duties are those requirements that apply to persons qua persons; they are not grounded in any specific act or owed to any specific party. All humans may have a duty, for instance, to provide help to those in need; this requirement seems not to require any generative act, it merely applies to us as persons. Similarly, it seems not to be owed to any particular person or persons, nor can it be "discharged" – "disposed of once and for all."[10] An obligation, on the other hand, is more personal in nature; it arises as the result of a voluntary act, and is owed to a particular party. It can also last for a limited time; for as long as the required act takes to complete, for instance.

Consider the following example. My neighbour helps to clean my pool, an act for which I promise to pay him £10. Promissory acts are, in my view, quite uncontroversial cases of obligation creation; a voluntary act (the utterance of the promise), binds one to act in a certain way (pay the neighbour £10). Furthermore, the obligation is owed by a specific person (the promisor) to a specific person (the promisee). The moral force of an obligation, then, arises through the voluntary and personal nature of certain interactions. This voluntaristic view is an appealing one, in that it acknowledges the importance of our autonomy at the heart of our moral landscape; obligations are morally significant because they are willingly generated by autonomous individuals, and not simply forced upon agents.

Content independence is another important feature of obligations. The binding nature of obligations arises not through the specific nature of the required act, but through the type of voluntary interaction characterised in conditions (1) and (2). I am obliged to pay my neighbour £10 because I promised to do so, not because the act of giving £10 to one's neighbours is generally a good thing. Contrast the example with one in which my neighbour is in dire need of money to secure the basic means of survival, but received no promise from me

[9] Hart, 'Are There Any Natural Rights?' in *The Philosophical Review*, Vol. 64, No. 2 (Apr., 1955), p179.
[10] Simmons, *Moral Principles and Political Obligations* (Princeton University Press, 1979), p14.

regarding any payment. In this case, I may have a *duty* to help him; one grounded in a general moral duty to help those in need. But this duty would (a) not have arisen out of a voluntary act, (b) not be owed to my neighbour in particular,[11] and thus would not be an obligation.[12]

Condition (4) links rights to obligation. Because obligation emerges in the kind of voluntary relationships characterised in features (1) and (2), binding obligations confer rights upon the specific parties to whom they are owed. If I am obligated to pay my neighbour £10, for instance, then that neighbour has a right to such payment. This is what Simmons terms the "logical correlativity of rights and obligations"; the existence of the latter entails the existence of the former.[13] We may note, however, that the reverse isn't necessarily true; we have many rights that are not grounded in voluntary obligation-creating acts, such as the right to privacy. Such rights correlate not with obligations, but with natural duties; those which, as noted above, simply apply to persons qua persons. This point turns out to be an important one, and shall be revisited in the next section.

The features above combine to provide a lucid picture of state legitimacy. Legitimate authority possesses some right to rule; something that distinguishes it from mafia bosses and other such powers. On the picture that emerges from Hart's conditions, such a right is easily understood; the state has a right to rule insofar (and to the extent that) its putative subjects have, through their voluntary acts, obligated themselves to obey it. Such obligations are owed to a particular body (the state), and are content-independent; one is obliged to obey the state's directives regardless of the nature of the actions they require.[14] The undertaking of such obligation simultaneously confers rights on the state; the right to issue directives, and enforce them – in short, the right to rule.

Evidently, this is only one of many possible conceptions of legitimacy. One may, for instance, take legitimacy to require the equal treatment of citizens, the acceptance of some proportion of individuals within a territory, or

[11] In addition, it would be unlikely to specify a particular sum of money that it would be appropriate to contribute.
[12] This isn't, of course, to say that obligations are somehow stronger requirements than duties. It is plausible to say, I think, that duties will sometimes take precedent over obligations (when, for instance, one has to break one's promise to meet a friend at the cinema in order to help a drowning child), and vice versa.
[13] Ibid, p14-15, 195.
[14] This assertion may, in fact, be too strong. It would be at least plausible to posit that obligations to perform immoral acts, such as murder, cannot be binding.

recognition in the international community.[15] Such features, I think (particularly the former two) are relevant to the moral assessment of states. In my view, however, they should not be understood as factors bearing on the legitimation of the state; since none immediately seems to ground the special kind of moral relationship necessary for a state to have the illusive right to rule over individuals. Such considerations may, however, be invoked in a *justification* of the state; a concept to which I return later in the chapter.

It should be noted that condition (4) above – the logical correlativity thesis – is not accepted by all; some (M.B.E. Smith and Rolf Sartorius, for instance) argue that legitimacy and obligation should be kept separate.[16] Holders of such a view may argue that political authorities can be legitimate even if individuals are not bound to follow their directives.

On the appealing Hartian framework just proposed, this position has no place; state legitimacy and political obligation are simply two sides of the same coin. The view does, however, seem to capture an intuitive thought; the idea that we can have something to say about the moral status of authorities quite apart from the question of obligation. This thought is better expressed, I venture, in terms of a distinction between legitimacy and justification. Briefly stated, the point is that political authorities can be *justified* even if they are illegitimate (and thus, even if individuals are not bound to follow their directives).

Before fleshing out the content of this distinction, allow me to provide further motivation for it. The merit of such a distinction appears most pressing, I think, in light of the challenge presented by philosophical anarchism, according to which no states (or very few states) are legitimate.

Arguments for the legitimacy of states, and philosophical anarchism

Attempts to show that states are (or a given state is) legitimate have been legion, and have taken many forms. The task, if we characterise legitimacy along the lines suggested so far, is one of showing that the putative subjects of the state

[15] Simmons, 'Justification and Legitimacy,' in *Ethics*, Vol. 109, No. 4 (July 1999), p747-748.
[16] See Smith, 'Is There a Prima Facie Obligation to Obey the Law?' in *The Yale Law Journal* Vol. 82, No. 5 (Apr., 1973), p976; and Sartorius, 'Political Authority and Political Obligation,' in *Virginia Law Review* Vol. 67, No. 1, The Symposium in Honor of A. D. Woozley: Law and Obedience (Feb., 1981), p4.

have, by some voluntary act (or omission), obligated themselves to obey the state. This, given the logical correlativity of obligations and rights, would confer legitimacy on a state.

Evidently, the picture of legitimacy presented thus far is such that authorities can be legitimate relative to each individual; if I voluntarily obligate myself to obey a body, that body has legitimate authority over me. The legitimacy claim of the state, however, is more far-reaching than this. States do not (or at least, have rarely been known to) allow for the possibility that some individuals within the territory over which it claims a right to rule are not bound to obey it. The directives and laws issued by the state are intended to apply to everyone within its purported domain; those who deny its legitimacy relative to themselves are subject to the same coercive enforcement as those who voluntarily obligated themselves to obey.

States, then, usually claim universal legitimacy within a given territory. Relatedly, many accounts of legitimacy appear to assume (explicitly or otherwise) that the task at hand is one of showing how *everyone* in a territory (or even, all territories) is bound to obey the state. This requirement of unanimity, however, seems to be overly restrictive, obscuring the nature of obligation. If obligation is taken to be voluntaristic (as I think that it should be), the possibility that some, and not others, are obligated to obey the state should not be considered problematic.[17]

What, though, of the semantic issue concerning the word 'legitimate?' It has been assumed so far that an authority can either be (wholly) legitimate or (wholly) illegitimate with respect to each individual. How, though, should we apply the term when speaking about the state in relation to a population or group? Should we thus only use the term 'legitimate' to describe those states in which *every* individual is obligated to obey? Or could we use it to describe a state to which, say, 51% of people within a territory are obligated?

Note that, by raising the question, we acknowledge that states can be more or less legitimate with respect to entire populations; one to which 100% of individuals voluntarily obligate themselves would have greater legitimacy than one to which only 70% do. Whether, then, one wishes to reserve the term

[17] This is true at least in a philosophical sense. Those who are not obligated to obey a state may find the fact that they are coerced by it anyway to be problematic in quite another manner.

'legitimate' to those authorities that achieve unanimous obligation, is a matter of preference.[18] I shall not employ that restriction, but shall also stop short of defining the exact threshold at which an authority may be said to be legitimate; following Simmons, I shall (very roughly) require that political obligation be "general"; that "most (or at least many) citizens […] are politically bound."[19]

Another requirement employed by Simmons also seems pertinent. An account of political obligation, claims Simmons, must be such that it explains how individuals are bound to *their* state in particular, and not just any group of states that fulfils certain criteria. This "particularity requirement" draws on the distinction between obligation and duty; an individual may have, for instance, a duty to support and promote just states, but this does not bind that individual "to one *particular* state above all others, namely that state in which he is a citizen."[20] On this view, non-particularised duties, given that they do not arise from a morally significant transaction between specific parties, do not amount to political obligations of the kind that entail the legitimacy of states.

The requirement of particularity also relates to the demands that our own particular state places upon us. It is plausible to think we owe support to all just states (or at least that we owe them *something* in a way we don't owe unjust ones), but the states that we are born into *demand* my support in a unique manner; they seek to impose obligations, and back up their demands with the threat of coercion. Added to this, a state may even make demands on individuals who subsequently leave that state's purported territory; by requiring that I continue to pay taxes whilst working in a foreign country for instance. Such observations make the particularity requirement a useful one for tackling states on their own terms.

The requirements of generality and particularity form the basis of Simmons' version of philosophical anarchism. Briefly stated, philosophical anarchism is the denial of political obligation (and hence, the denial of the legitimate authority of some or all states). The position is a relevant one, because it highlights the need for tools for morally assessing the state that go beyond the

[18] One could, I assume, use the phrases 'more legitimate' and 'less legitimate,' to assess those authorities to which there is not unanimous obligation, even if one desired to reserve the adjective 'legitimate' for the unanimous cases. Consider the comparison of two groups of people, one of which is collectively whispering, one of which is silent. One might describe the whispering group as 'louder,' even if neither group is 'loud.'
[19] Simmons, *Moral Principles and Political Obligations* (Princeton University Press, 1979), p55.
[20] Ibid, p32.

issue of legitimacy. If the philosophical anarchist stance is a plausible one – that is, if it is possible that no states are legitimate – then, in the absence of any further means of evaluation, we might take them all to be on an equal normative footing. This thought, I shall suggest, is a troubling one, and one which motivates the employment of the kind of *justificatory* assessments that form the core of this paper.

The sceptical philosophical anarchist position may take different forms. Robert Paul Wolff, for instance, attempted to provide an "a priori argument" for the impossibility of a legitimate state; he claimed that authority, insofar as it inherently conflicts with personal autonomy (the preservation of which is a "primary obligation"), could never be legitimate.[21] Wolff's position, however, is a problematic one. Firstly, he admits that "there are at least some situations in which it is reasonable" to sacrifice a part of our autonomy; when we visit the dentist, for example.[22] It is not immediately obvious why the undertaking of political obligations could never constitute such a situation. Secondly, Wolff actually unearths a political system – unanimous direct democracy – that, if instituted, would resolve the conflict between authority and autonomy.[23] That such a system would be difficult to maintain does not hide the fact that legitimate authority, on Wolff's own terms, is at least possible in theory.

In my view, a more plausible version of philosophical anarchism is presented by Simmons. In contrast to Wolff's view, Simmons' project is decidedly "a posteriori" in nature; he nowhere denies the possibility that states could be legitimate, holding merely that, empirically, most people are not obligated to the states that coerce them. His position is a critical one; he explores, and subsequently rejects, a range of ways in which it has been argued that states are legitimate.

A full exegesis of the philosophical anarchism proposed by Simmons is evidently beyond my remit here. It will suffice, for my purposes, to sketch the main thrust of the position, such that its plausibility begins to emerge. I find Simmons' account to be an appealing one, but a full endorsement is neither

[21] Wolff, *In Defence of Anarchism* (University of California Press, 1998), p19.
[22] Ibid, p15.
[23] Ibid, p22-27

offered, nor necessary. The point here is simply to note the challenge posed by the philosophical anarchist to the would-be assessor of the state.

Simmons roughly divides the contesting accounts of political obligation into three sets; "transactional," "associative," and "natural duty" stories.[24] Transactional accounts seek to ground political obligation in some specific voluntary undertaking with the state; the kind of transaction, we may assume, that is suggested by the Hartian characterisation of obligation itself. Associative accounts, meanwhile, make reference to the (nonvoluntary) roles and positions of individuals within a certain political framework, whilst natural duty theory invokes duties that we have qua persons.[25]

The associative group falls down, on Simmons' view, primarily due to a lack of voluntarism and generality. The principle of fair play perhaps offers the most feasible version of such an account. The principle, in the words of Rawls, "holds that a person is required to do his part as defined by the rules of an institution when two conditions are met: first, the institution is just […]; and second, one has voluntarily accepted the benefits of the arrangement…"[26]. Initial formulations of this view, (such as that of Hart) suggested that we are politically obligated merely due mere 'receipt' of benefits. This would appear to bind us too easily to states; it seems to imply that a cooperative scheme may be forced upon us, as in Nozick's famous public address system and book giving examples.[27] Appeal to 'acceptance' of benefits may rule out these counterintuitive cases, but seems to drastically limit the principle; since it seems implausible to suggest that most citizens have voluntarily accepted state benefits. Many will not have taken benefits willingly (and besides, the fact that we are born into political societies makes it nearly impossible to avoid them), as Rawls later admits.[28] The principle, then, seems not to ground the general voluntaristic undertaking of obligation that I have taken legitimacy to involve.

Natural duty theory is also dismissed by Simmons, on the basis that it fails to generate particularised obligation. A duty to support just states, as suggested above, doesn't give us the kind of specific political obligations that

[24] Simmons, 'The Duty to Obey and Our Natural Moral Duties,' in Wellman & Simmons, *Is There a Duty to Obey the Law?* (Cambridge University Press, 2005), p102.
[25] Ibid, p109.
[26] Rawls, *A Theory of Justice: Revised Edition*, (Belknap, 1999), p96.
[27] Nozick, *Anarchy, State and Utopia* (Blackwell, 1974), p94.
[28] Rawls, *A Theory of Justice: Revised Edition*, (Belknap, 1999), p296.

states demand of us. One may try to supplement the account, suggesting that I owe a duty of support only to the institutions that 'apply to me'; namely those of the state which claims authority over me. But it is hard to see how any obligation can stem from this – the fact that I was born in a certain place rather than elsewhere is morally arbitrary. To have any force, this theory would have to claim a second, stronger sense of application, one that seems likely to refer to some kind of transaction with the state.

Transactional accounts seem to offer the best hope for an account of political obligation. One version in particular – consent theory – has long occupied pride of place in debates in political philosophy, and for good reason. In the same way as promising was taken to be an uncontroversial way of generating obligation, so consent (if it is freely and knowingly given) confers the correlating rights and obligations on the parties involved. When I voluntarily consent to something, I do so with the intent of altering my moral landscape; I undertake to create a binding agreement between myself and another specific party. Consenting to the existence of the state, then, involves granting it a right to rule, and thereby generates particularised political obligation. Simmons, it seems to me, takes it to be uncontroversial that unanimous consensual agreement among individuals would legitimate a state.

The problem, however, is that most of us have plainly not offered actual consent to the state. As Simmons notes, "Real citizens in real political communities seldom do anything that can be plausibly described as either a promise to obey or any other kind of freely made commitment to comply with domestic laws."[29] Whilst actual consent would legitimate the state, then, there is a problem of generality. Although some individuals (naturalised citizens, for example) may have expressly agreed to acquire political obligations, it is simply not the case that most (or even many) of us have done so. Accounts that invoke tacit consent, meanwhile (such as Locke's claim that residence signifies consent), whilst providing the necessary generality, seem to fall down in terms of voluntariness; since the choice we face (with regards to residence at least) only really amounts to a choice *between* states, all of which make similar demands upon us.

[29] Simmons, 'The Duty to Obey and Our Natural Moral Duties,' in Wellman & Simmons, *Is There a Duty to Obey the Law?* (Cambridge University Press, 2005), p118.

Simmons' position, then, (unlike that of Wolff) allows for the fact that some may have political obligations. It simply posits that the rarity of such cases undermines the claims of most states to a generalised legitimacy over their putative subjects.

The pessimistic view of state legitimacy presented by the philosophical anarchist, then, may be thought to pose a problem, in two respects. Firstly, we initially seem to be denied the conceptual tools to normatively distinguish between different states; the philosophical anarchist's verdict is simply that all states are illegitimate. This thought is an uncomfortable one, since we commonly tend to think of some states as preferable to others.

Secondly, if all states are illegitimate, their coercive nature appears objectionable to an extent that may warrant significant practical remedy. If most people obligated themselves to a state, then that state may have the legitimate authority to impose duties upon individuals and enforce them. Without that legitimation, goes the argument, state coercion amounts to little more than large-scale bullying; behaviour that it would be sensible to abolish. Most people, however, including philosophical anarchists themselves, do not subscribe to a more political anarchism. This may, to some extent, be down to a fundamental lack of understanding of the nature of the state itself.[30] Another possibility, however, is that many people do recognise the objectionableness of being coerced, yet for whatever reason accept the situation.

This second suggestion is a plausible one, and leads into a discussion of distinction between legitimacy and justification. These, I hold, are two different ways of morally assessing the state. Even when a state is judged to be illegitimate, then, we may still say something about its normative status, in such a way that accommodates the intuition that some states are preferable to others.

Justification and legitimacy

Legitimacy, as defined thus far, is conferred on a state by voluntary obligation-creating undertakings by the individuals over whom it claims a right

[30] Simmons, for instance ventures that "Beliefs about political obligation – insofar as we actually have any – are […] "suspect" as a kind of "false consciousness" that it serves the interests of powerful others to induce in us." Simmons, 'Philosophical Anarchism,' in Sanders/Narveson ed. *For and Against the State* (Rowman and Littlefield, 1996), p33.

to rule. States, however, even if we take them – as the philosophical anarchist does – to be illegitimate in this way, may still be more or less desirable; they may exhibit features or perform certain functions that make their existence a good thing. This thought forms the basis of a second way of assessing states; that of justification.

What is it to justify something? When I justify, for instance, my actions, what I am I doing? The first thing to note is that an act of justification (whether of actions, practices, institutions, or anything else) usually seems to be demanded or offered in light of some potential objection, whether explicit or otherwise. I might need to justify, for instance, a decision that imposes costs on my neighbours, in a way that doesn't seem necessary for a decision that does not. This is what leads Simmons to characterise justification as a "defensive" concept; he notes that "we ask for justifications against a background assumption of possible objection."[31]

In the case of the state, we may posit, that presumption takes the form of a prima facie objection to coercion. As autonomous decision-makers, we hold a presumption against having actions imposed on us from without, and against being threatened with violence.[32] This, I take it, is our objection to mafia bosses, and those who threaten us in the street; they impinge upon our liberty in a manner we find unacceptable. Of course, one way in which we might find it acceptable to be coerced involves legitimacy; a state, for instance, to which everyone in a territory has voluntarily obligated themselves through express consent, would hold a right to rule that renders the background presumption against coercion unimportant.

Alternatively – and even when it is illegitimate – a coercive state may be justified. Justification typically takes the form of showing (in the face of the kind of background prima facie objection noted above) that an act or phenomenon is "prudentially rational, morally permissible, or both."[33] I may justify my imposition of costs on my neighbours, for instance, by pointing to the fact that, in so doing, I prevented them from incurring even greater costs. Justifying the state, then, would involve demonstrating that it "is on balance morally permissible (or

[31] Simmons, 'Justification and Legitimacy,' in *Ethics*, Vol. 109, No. 4 (July 1999), p740.
[32] Note that here I speak only of a 'presumption' against coercion. I do not claim, as Wolff does, that we have any kind of overriding duty to preserve our autonomy.
[33] Ibid, p740.

ideal) and that it is rationally preferable to all feasible nonstate alternatives."[34] In the words of Robert Nozick, to justify the state would be to show it to be "superior to even [the] most favoured situation of anarchy."[35]

The appeal to prudential considerations here may take many forms. The state may be justified insofar as it exhibits certain characteristics, furnishes individuals with specific benefits, and so on. In the following chapter, I explore what I take to be the most compelling version of such a justificatory story; that which invokes the shortcomings of individual rationality. I do not claim that this is the only plausible manner of providing justification for the coercive state, and the enterprise may still be a fruitful one even if, as I argue, this particular version of it fails to convince.

Justification provides a metric by which to normatively assess states and their coercive practices, regardless of whether they are legitimate or not. The presence or absence of binding obligations upon individuals needn't settle the further question of whether a state is justified. A state to which, for instance, we have not consented, can still be justified if we find its existence to be morally permissible and to constitute, in whatever way we specify, a "good thing," rationally speaking.[36] In such cases, the practical concern over a more political anarchism is allayed; even if we are not morally obligated to the state, we may have reason to accept, comply with, or even support states whose existence we find to be justified.

It has been posited that the state, even if illegitimate, can be justified. Does the need for justification arise, though, in the case of legitimate states? So far, it has been suggested that the prima facie objection to coercion might be overcome when individuals voluntarily obligate themselves to the body doing the coercing. A legitimate state's coercive acts may be thought not to require justification; it has a right to rule, which removes the background objection to coercion. What though, of the possibility that a legitimate state (one, for ease of

[34] Ibid, p742.
[35] Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974), p4-5.
[36] The distinction between legitimacy and justification may be stated in different terms. David Schmidtz, for instance, delineates between "emergent" and "teleological" forms of justification. The former equates roughly to what I am calling legitimacy, and involves focusing on "the process by which the state comes to be." Teleological justification, meanwhile, is closer (but not necessarily analogous) to what I am calling justification: a form of assessment that looks primarily at "what [states] accomplish." Schmidtz, 'Justifying the State,' in Sanders/Narveson ed. *For and Against the State* (Rowman and Littlefield, 1996), p82.

argument, to which all individuals in a territory have explicitly consented) perform badly, such that its existence is not a good bargain for its citizens?

Such a case underlines the fact that justification and legitimacy, although distinct, may be in some sense complimentary in terms of the moral assessment of the state. Firstly, it would seem irrational for individuals to voluntarily obligate themselves to a state which they believe will fail to fulfil the functions and exhibit the characteristics that they desire in a state. The very legitimation of a state, then, might involve some prior consideration of the factors that play a part in justification.

Secondly, the possibility that a legitimate state become prudentially unjustifiable over time highlights something important about the type of moral assessment in question; the fact that our relationship with the state may be more sensitive to temporal shifts than a theory of political obligation can account for.[37] Imagine, for example, everyone in a given territory consents to the state, on the condition that it will perform certain functions. Imagine, subsequently, that the state fails in those respects (and/or acts in such a way such that its existence is not in the interests of individuals). It may strike us as problematic that a single set of voluntary undertakings can legitimate the state in perpetuity; that we may be obligated to a body that may be constantly changing in nature and efficacy.[38] Given the (practical, if not theoretical) difficulty of a state having to constantly legitimate itself, though, the justificatory assessment may be the most salient practical tool for evaluating the state.

[37] A clarification is needed here. I take it that a justification is an all-or-nothing concept with respect to individuals; that is to say that for each agent, the state either is or is not justified. Evidently, one state may represent a better deal than another; but if both provide some certain level of benefit (relative to the individual in question), they may both be justified. At the level of society, however, the term 'justification' may be taken to admit degrees, corresponding to the proportion of individuals within a putative territory who take the relevant authority to be justified. Note that this corresponds to my use of the term "legitimate" earlier in this chapter.

[38] A further problem would be caused by the state acting beyond its remit: i.e. involving itself in domains that the consenting parties did not agree to.

CHAPTER 2 – GAME THEORETICAL JUSTIFICATION

The task of justifying the state, then, is one of providing reasons for the acceptance of its coercive acts; something which we normally take to be prima facie morally problematic. Such accounts may take many forms, but I focus on one main argumentative strategy, which has been taken by many to constitute the most convincing argument that the state is justified.[39]

That position, briefly stated, takes the following form. Acting on individual rationality (defined in a certain way) alone, agents cannot furnish themselves with certain goods that are assumed to be of great importance to them. Through its coercive function, however, the state can ensure the behaviour that leads to the generation of such goods. The state, on such a view, is justified insofar as it facilitates the production of such goods; if it fulfils certain functions in this way, individuals have reason to accept the imposition of state coercion even if that state is not legitimate.

Such arguments frequently invoke a game theoretical framework. By defining rationality in a specific way, and by considering interactions between individuals as strategic 'games,' the would-be justifier of the state seems, as we shall examine, to be presented with a plausible theoretical basis for their position. One type of game in particular has received a great deal of attention within the literature; namely the prisoners' dilemma. I shall briefly discuss the game theoretical enterprise in general, and the prisoners' dilemma in particular, before presenting the justification of the state of which they form the foundations.

Prisoners and the putative shortcomings of individual rationality

Two individuals, Anne and Bill, have been arrested on suspicion of committing a robbery. They are held in separate cells, and as such have no means of communicating with each other. Anne and Bill are questioned by the local deputy of police, who informs them that, as things stand, he has insufficient evidence for a robbery conviction. He does, however, have the evidence to

[39] I take it that the argument that follows is intended, by at least some of its proponents, to be sufficient for justification. I do, however, leave open the possibility that it may – if successful – be employed as part of a wider justificatory project.

prosecute for a lesser offence. Each prisoner has the same two options; either remaining silent, or agree to testify against the other. Let us label the former option as (C), and the latter option as (D).

The deputy informs each of the detainees that their sentence will depend on the choice (between (C) and (D)) that he makes, and on the choice made by their putative partner in crime. There are then, four possible ways in which the agents' choices can combine; CC, DD, DC, and CD. If each agent testifies against the other (DD), each will be convicted of robbery, and handed a sentence of 5 years in prison. If, however, they both remain silent (CC), they will each be handed a lesser sentence of 2 years, due to lack of evidence on the count of robbery. If one testifies to their partner's guilt, and the other stays quiet ((DC) or (CD)), then whoever stays quiet will be sentenced to 10 years, whilst the defector will get off scot-free. The sentences for each of these combinations are listed in the table below;

Anne's sentence (years)	Outcome	Bill's sentence (years)
0	DC	10
2	CC	2
5	DD	5
10	CD	0

We may assume in this situation that both Anne and Bill prefer to spend fewer years in jail. It is also presupposed that neither is influenced by the potential sentence of the other. The sentence matrix above, then, can be represented in the following matrix, which lists the preferences of each individual in an ordinal manner (that is to say, for each agent, preferences conform to the following ordering; $1 > 2 > 3 > 4$);

Anne's preference schedule	Outcome	Bill's preference schedule
1	DC	4
2	CC	2
3	DD	3
4	CD	1

With this much established, we may draw up one last table; one in which both Anne and Bill's ordinal preferences are represented in a 2 by 2 matrix. In each square, the first figure represents Anne's preference, the latter figure Bill's.

		Bill	
		C	D
Anne	C	2, 2	4, 1
	D	1, 4	3, 3

What, then, should each individual choose to do? With prison looming, should they stay quiet (C), or agree to testify (D)? What will the outcome be? The above matrix seems to provide answers to these questions. Consider first the point of view of Anne. Looking at the table, she takes into account only the first number in each box, and wishes to secure an outcome that satisfies her preferences as much as possible. She notes that, should Bill decide to stay quiet (C), she would do better by defecting (D); since, evidently, she prefers her first best preference to her second best. She also observes that, if Bill decides to defect (D), she does better by defecting as well; since this would secure her third preference instead of leaving her with her least-favourite option. Anne, then, realises that, *whatever* Bill does, she does better by choosing to incriminate him. Now consider Bill's situation. Looking at only the second numbers, Bill notes that, should Anne choose (C), he does better by opting for (D). If Anne chooses (D), on the other hand, then (D) is also his best response. His reasoning, then, is symmetrical to Anne's; whatever she does, his preferences are best fulfilled by agreeing to testify against her.

On the basis of such reasoning, then, both Anne and Bill testify against one another. As a consequence, outcome (DD) comes about, and each of them is handed a prison sentence of five years. The deputy of police, having witnessed all of this first hand, chuckles. "These shmucks never learn," he whispers to himself; "if they had both stayed quiet, they would have only got a couple of years each!"

The story above is a version of a situation first put forward by A. W. Tucker. The tale serves to roughly characterise a much-discussed analytical tool within philosophy, as well as in many other disciplines; the prisoners' dilemma.

Let us now highlight the more formal elements of the situation, locating it within a game theoretical framework.

Firstly, the interaction between the two parties is characterised as a 'game'; one in which the combination of their choices (or *strategies*) determines the *payoffs* that they will receive. In the case above, the payoffs come in the form of prison sentences; each strategy vector (combination of the strategies of the two players) has an associated payoff to each player. Such games, as shall be mentioned later, may be modelled in a manner that involves any number of players. For the instant, however, it will suffice to confine our attention to the two-player game.

Secondly, game theory appeals only to rationality as the basis of choice; each player is to decide upon a strategy in light of a rational deliberation of his situation. Such rationality, however, is defined minimally; rationality in the theory of games is assumed to be individual and maximising.[40] Players pursue the maximal satisfaction of their own preferences (and thus seek to maximise their own payoff in the game), without regard for the payoffs afforded to other agents. This strictly defined individual rationality may initially seem somewhat detached from everyday human experience (a point I shall revisit later), but it does, at first glance at least, allow game theoretical models to provide clear, lucid explanations of certain behavioural phenomena.

The third feature of the prisoners' dilemma is that each of the players has a dominant, non-cooperative strategy. In the story above, the strategy of testifying against the other strictly dominates the strategy of staying quiet for each of the agents; it offers a better payoff no matter what the other player decides to do. Why, though, are the dominant strategies said to be non-cooperative? As we shall see, the (real world) situations to which the model is applied are usually such that the abbreviations (C) and (D) stand for cooperate and defect; the strategy D, then, and the outcome DD, are usually described as non-cooperative. I shall generally employ this terminology throughout this paper,

[40] The exact specification of rationality that is employed (or should be employed) within such models is a subject of debate. Sen, for example, distinguishes between three distinct ways in which individual maximising rationality can be egoistic or self-interested. See; Sen, *Rationality and Freedom* (Belknap: Harvard, 2002), p33-34. For my purposes, I shall primitively take the assumption at the heart of game theory to be one of individualistic maximisation; that rationality involves simply that one seek to maximise the satisfaction of one's own preferences, no matter what their source.

although it should be noted that it is the specific preference structure, rather than any specific types of actions, that defines the prisoners' dilemma. Any game or situation in which preferences are those detailed above may be said to be a prisoners' dilemma.

In the story above, Anne and Bill were held in separate cells, as to prevent communication between them. Without the means to contact one another, the thought is, there is no way for them to make an arrangement whereby both agree to (C) and get 2-year sentences. We may note, however, that an assumption of non-communication isn't strictly necessary to the prisoners' dilemma. This is because, given the nature of the rationality that players are assumed to have, even pre-game agreements are unlikely to prevent non-cooperative outcomes.

Consider the following case. Anne and Bill, just after being told about the potential sentences they face, manage to communicate, and make a verbal agreement to each remain silent. Given the demands of maximising rationality, however, each player has reason to renege on their agreement. Anne sees that, should Bill stick to the agreement, she does better by defecting; she'll be let off scot-free. Likewise, Bill is also swayed by the possibility of betraying Anne; the *temptation* between a 2-year sentence and no sentence means that he too will choose D. The outcome, then, is exactly as it would have been had there been no agreement whatsoever.

The example demonstrates that, within the prisoners' dilemma framework, pre-game agreements between parties are not binding. These are situations in which, as the saying goes, *talk is cheap*; promises and the like simply hold no sway because of the kind of simple maximising rationality at play. It may, to be sure, be rational to enter into such pre-game agreements – doing so opens up the possibility of exploiting the other player – but when the moment for action arises, the rational thing for a player to do is renege on the deal.

Finally, we must introduce the concept of Nash equilibrium. The Nash equilibrium in a two-player game is the point at which "for each player, the

strategy choice selected is a best reply to the strategy choice of the other."[41] If, then, there is a strategy pairing from which it pays neither individual to deviate to another available strategy, that pairing is said to be a Nash equilibrium within the game. In the prisoners' dilemma, then, the strategy pairing DD is evidently such an equilibrium; if either party decided to switch to a C strategy, they would do worse than if they had maintained their D strategy. The Nash equilibrium point in the prisoners' dilemma is the non-cooperative pairing DD.

The above features, then, provide a brief characterisation of the prisoners' dilemma. One obvious question, however, remains to be expressly answered; why is the prisoners' dilemma a dilemma? Simply put, games of this type are taken to spell out shortcomings in individual rationality. Recall that the Nash equilibrium point DD gives each player his third preference. Both players, to be sure, prefer this outcome than the one in which they are exploited by the other player. But within the 2 by 2 matrix, there is another strategy pairing that *both players* would prefer to DD; the pairing in which each player chooses C.

Using the terminology employed by game theorists, we may say that the strategy vector CC is *Pareto-preferred* to the vector DD. One outcome is described as Pareto-preferred to another if and only if "at least one player (strictly) prefers the first to the second and no player (strictly) prefers the second to the first."[42] In the prisoners' dilemma, both parties strictly prefer CC to DD; since each of them prefers to obtain their second-best preference to their third-best.

The prisoners' dilemma, then, highlights a situation in which agents seeking to maximise the satisfaction of their preferences end up with payoffs that are Pareto-inferior to those offered by another available outcome. Individual rationality, in other words, is such that agents cannot obtain the advantages that would arise through truly cooperative interaction. The decisions of rational agents alone cannot bring about behaviour that is mutually beneficial.[43]

---

[41] Harriott, 'Games, Anarchy, and the Nonnecessity of the State,' in Sanders/Narveson ed. *For and Against the State* (Rowman and Littlefield, 1996), p124.
[42] Taylor, *The Possibility of Cooperation* (Cambridge University Press, 1987), p63.
[43] One may wish to employ the term "mutually rational" to describe the outcome that eludes individuals in the prisoners' dilemma. This phrase, however, seems likely to confuse matters, given that I have frequently stressed the individual nature of the rationality employed in the framework. I shall, therefore, use the term "mutually beneficial."

Public goods and the economic justification of the state

The prisoners' dilemma, and its pessimistic message about individual rationality, has widely been taken to form the basis of attempts to justify of the coercive state. In particular, it is at the centre of what has often been labelled the "economic argument for the state."[44] According to such a view, the state is justified insofar as it solves (or rather, prevents agents from finding themselves in) prisoners' dilemma situations within certain real-life interactive situations; those which concern the provision of public goods.[45]

The notion of a public good, so frequently employed in economic contexts, must be clarified here. Following Michael Taylor, I will use the term *public good* to denote a good (or service) that is "in some degree indivisible and non-excludable."[46]

A good or service is perfectly indivisible if, "once produced, any unit can be made available to every member of the public, or equivalently, if any individual's consumption or use of the good does not reduce the amount available to others."[47] In contrast to private goods, then, units of public goods cannot be appropriated or used up to the detriment of other individuals. A chocolate bar, for instance, is perfectly divisible; my consumption of it reduces the amount of chocolate bars available for consumption by others by one unit. By contrast, my use of well-maintained road system does not (normally) subtract from the total amount of the good available to others.

As Taylor notes, indivisibility should not be confused with non-rivalness, a notion occasionally employed in this context. A rival good is such that one individuals consumption reduces not the amount available to others, but the "*benefits* to others who consume that same unit."[48] Evidently, perfectly divisible goods will also display a high degree of rivalness (my eating of a chocolate bar

[44] Harriott, 'Games, Anarchy, and the Nonnecessity of the State,' in Sanders/Narveson ed. *For and Against the State* (Rowman and Littlefield, 1996), p120.
[45] The prisoners' dilemma has also been taken to represent certain other situations that might be argued to have a bearing on our view of the state. The likes of Gregory Kavka, Jean Hampton, and David Gauthier, for instance, have all written extensively about the possibility (or otherwise) that Thomas Hobbes' views on the state of nature be adequately mapped by the kind of game theoretic tools in question. This question is undoubtedly an interesting one, but would seem to require substantially more attention than I may grant it here. I therefore restrict my attention to the public goods application of the prisoners' dilemma.
[46] Taylor, *The Possibility of Cooperation* (Cambridge University Press, 1987), p5.
[47] Ibid, p6.
[48] Ibid, p7.

reduces both the amount and the benefits available to others), but the two notions can come apart. Taylor elaborates;

> An individual's benefit from consumption may not change at all as the amount available for consumption declines… in fact, some individuals' utilities may rise as the number of other customers increases, at least up to a point; they may, for example, prefer a semi-crowded beach or park to an empty one.[49]

This example, claims Taylor, highlights the fact that non-rivalness, unlike indivisibility, seems to pick out a property of individuals' utility functions, rather than a feature of public goods. I shall employ the notion of indivisibility rather than that of non-rivalness for this reason.

We may describe a good or service as non-excludable, meanwhile, if individuals (within a certain group or population) cannot be prevented from consuming it, or if such prevention would come at an unreasonably high cost.[50] Non-excludability, then, implies open access to a good once it is produced or secured. Clean seawater would be an obvious example of a non-excludable good; it would be very difficult indeed to prevent individuals from benefitting from a policy that aimed at securing such a good.

The nature of public goods, then, outlined in the two features above, allows for the possibility of *free-riding*; consuming a good without having contributed to its production. Given that individuals cannot be prevented from accessing the goods in question, and that their use of the goods will not reduce the quantity available for others, agents are presented with the chance of exploiting the efforts of others.

This possibility has led many to posit that the preferences of individuals in public goods interactions (i.e. situations in which agents must decide whether or not to contribute to the provision of a public good) are (or are often) the same as those found in the prisoners' dilemma game. In public goods interactions, players have two strategies available to them; to contribute to provision of the

[49] Ibid, p7.
[50] I take it that the question of exactly what cost would be reasonable or unreasonable to incur to secure exclusion needn't be resolved here.

public good (C), or to not contribute (D). All players, it is assumed, prefer there to be mutual contribution to production (CC) than for there to be no public good produced at all (DD). All players, however, would most prefer to benefit from the good without contributing than to benefit from it having done one's part. The good – if provided – affords a net benefit to each individual, regardless of whether or not they contributed to its provision. This is to say that they, as rational players in a game theoretical scenario, prefer to free ride when they can. Nobody, however, wants the sucker payoff; contributing to the provision of the good before everyone (including free-riders) reaps the benefits.

The preference schedule, then, is that of the prisoners' dilemma;[51]

Player 1	Outcome	Player 2
1	DC	4
2	CC	2
3	DD	3
4	CD	1

Which gives the following prisoners' dilemma matrix;

		Player 2	
		C	D
Player 1	C	2, 2	4, 1
	D	1, 4	3, 3

As before, then, strategy D is dominant for each player; each best satisfies his preferences by choosing not to contribute, no matter what the other decides to do. Individual rationality alone, then, produces the outcome DD; nobody ends up contributing to the production of the public good, and the good does not get produced.

This, of course, seems problematic; the kind of decision-making in question leaves individuals unable to attain benefits that would be within their reach if they managed to act cooperatively. The issue, it should be stressed, is not one of individuals simply failing to see that such public goods are beneficial to them. Rather, public goods are acknowledged by agents to be "highly desirable and

[51] Evidently, most public goods interactions will involve more than two players. The two-player game, however, is adequate for my current purposes.

essential for collective well-being."[52] The problem is simply that, with the possibility of free-riding such an attractive one, preferences are such that a collectively beneficial outcome is beyond the reach of agents behaving according to individual rationality. The pessimistic conclusion to the prisoners' dilemma, it seems, applies to public goods interactions.

The claim that such situations take the form of prisoners' dilemmas forms the basis of a justification of the state. If individual rationality alone cannot furnish agents with goods that – we may assume – are very important for their lives, then they may have strong reasons to accept some body or mechanism that does allow us to produce such goods. The state, it seems uncontroversial to posit, does provide (or at least, facilitates the provision of) such goods in many cases; education systems, transport links, and environmental are all, empirically, phenomena that we take to fall within the remit of the state.

The coercive nature of the state appears to be central to its ability to function in such a way. The state, through the threat of force, can attach substantial disutility to any strategy that would undermine an individual's contribution to the production of public goods. A simple example would involve tax evasion; states typically exercise coercive force to ensure that their putative subjects contribute monetarily to the funding of public goods.

Such altering of the prospective utilities attached to different strategies, then, can remove individuals from prisoners' dilemma situations. The available payoffs are assumed to be such that individual preferences are no longer those listed in the prisoners' dilemma; in particular, state enforcement reduces the rationality of any attempt to free-ride, since being caught would likely involve some penalty.

The precise payoff structure that arises from state enforcement will, of course, depend on at least two factors. Firstly, one must consider the severity of the punishment that attends a failure to contribute. Generally speaking, the greater the punishment (whatever specific form it may take), the more likely individuals are to be discouraged from running the risk of incurring it. Secondly,

[52] Harriott, 'Games, Anarchy, and the Nonnecessity of the State,' in Sanders/Narveson ed. *For and Against the State* (Rowman and Littlefield, 1996), p120. The true necessity of some such goods is, of course, open to debate, but it seems plausible to suggest that a certain a range of goods could be viewed as widely valuable among all (or at least most) individuals within a certain group.

and relatedly, one must consider the likelihood that an attempt to free ride will be noticed, and subsequently punished. In a state was somehow able to apprehend all such attempts, (and issue substantial punishments), then one would assume that this would render defection far less tempting than if it were only able to identify, say, 50% of attempts.

If the enforcement rate were high enough, and the disutility attached to sanctions significant enough, then an individual's preferences may be such that he prefers all outcomes in which he cooperates ((CC) or (CD)) to any that involve him playing a risky defection strategy. Thus, he may prefer (CD) – which is the 4th preference in the prisoners' dilemma – to mutual defection (DD), or even to (DC). In this manner, the state would ensure that mutually beneficial outcomes are secured.

The state, though, does not necessarily need a perfect ability to apprehend free riders in order to ensure public good provision. Securing sufficient cooperation may not require that individuals prefer exploitation (CD) to free-riding (DC), or even to mutual defection (DD). Perhaps the state's enforcement rate and punishments are only enough to render the outcome (CC) more preferable to free riding for each player; perhaps, for instance, the benefit of a successful free ride attempt over (CC) is outweighed by the risk of being caught. In this case, choosing to cooperate would be rational when one expected one's follow player to cooperate.

This last scenario would generate the following payoff structure;

Player 1	Outcome	Player 2
2	DC	4
1	CC	1
3	DD	3
4	CD	2

This game, then – one possible model of state enforcement of public good contribution – is not a prisoners' dilemma. Rather, it exhibits the preferences of different game;

		Player 2	
		C	D
Player 1	C	1, 1	4, 2
	D	2, 4	3, 3

D is no longer a dominant strategy for each player; rather, the state affords assurance to all players that the outcome CC will arise. It is in so doing – in its provision of a solution to the public goods prisoners' dilemma – that the state is taken to be justified. Even if a state is illegitimate, it may be rational to accept it (and its coercive acts) as a "pragmatic necessity"; since, "but for the state's activities to solve these public-goods problems, the goods will not be provided."[53]

Two points must be raised at this juncture. Firstly, and briefly, we must recognise the limitations of such an argument with regards to justifying the state. The account above makes reference only to one specific task (or set of tasks) that the state, empirically, tends to carry out. As such, it seems fair to say that such a position may only seek to justify the elements of the state (and use of coercion) relevant to that brief. It would, for instance, take a good deal more argumentative work to argue that all of the state's myriad directives and initiatives are justified by reference to its ability to solve public goods problems. It seems to me fairly doubtful that a plausible version of such an argument is possible, but I here leave open the possibility. I will, however, assume that the economic argument under consideration is intended only to justify those functions of the state that are pertinent to the provision of public goods. My opinion, as shall become clear, is that even such a justification of a minimal version of the state fails to convince.

Secondly, we must cast doubt on one of the assumptions employed in the economic argument. Whilst it seems uncontroversial to hold that states do (generally) provide a solution to the public goods problem, it would be a mistake, I think, to consider them the *only* possible solution. Empirical facts about the functioning of real states should not be taken as proof of any necessity.

Rather, as I shall argue in rest of the remainder of this paper, we should not be so pessimistic about the likelihood of individuals acting in a cooperative manner (and thus providing themselves with public goods) of their own accord.

[53] Ibid, p125.

The gloomy lesson that the would-be justifier of the state draws from the prisoners' dilemma, I argue, may be contested in two different ways. Firstly, it has been fruitfully argued that, even within the confines of a game theoretical framework, cooperative behaviour may be rational. I evaluate two such accounts in the following two chapters. Secondly, such accounts provide reasons for doubting the relevance of the prisoners' dilemma (and its pessimistic conclusion) within a normative assessment of the state.

CHAPTER 3 – ITERATING THE PRISONERS' DILEMMA

I will now consider the first of two projects that seem to undermine the notion that mutually beneficial cooperation cannot be achieved by rational agents in prisoners' dilemma scenarios. These accounts operate within the game theoretical framework, and aim to show that, even when agents' preferences at a given point in time are those described by the prisoners' dilemma, cooperative behaviour may arise in the absence of third-party enforcement.

The discussion in this chapter focuses on the introduction of temporal considerations to the prisoners' dilemma, and the supposed impact this has on the likelihood of cooperation. If the dilemma is recast as a series of games iterated over time, runs the argument, then new strategies may emerge; strategies that make cooperation far more feasible than in the one-shot version of the game. I will firstly consider the account of Robert Axelrod, whose employment of prisoners' dilemma computer tournaments provides an instructive introduction to the issues at hand. Axelrod's conclusions, however, receive a more persuasive philosophical treatment in the work of Michael Taylor, to whose project I subsequently turn.

Taylor, through the use of the notion of (Nash) equilibrium, seeks to demonstrate that, under certain conditions, rational agents may maximise their utility by using strategies that are conditionally cooperative. Taylor's position, however, is not unproblematic. I shall consider potential objections to his account, which reduce the force of his conclusion somewhat. I conclude that the iterative account does at least provide us reason to think that cooperation is more likely than it is taken to be in the one-shot analysis. Finally, I briefly suggest how certain considerations in this chapter will be expanded upon later in the thesis.

Supergames and strategy

In the one-shot prisoners' dilemma, players have only two strategies available to them; Cooperate (C) or Defect (D). As such, given that payoffs are a function of both players' strategic choice, there are only four possible outcomes; mutual cooperation (CC), mutual defection (DD), unilateral cooperation (CD) and unilateral defection (DC). As we have seen, D is the dominant strategy for

each player, which leads to the equilibrium point DD. The dilemma is that this outcome is, for each player, Pareto-inferior to the cooperative solution CC; individual rationality alone, it is concluded, cannot allow agents to realise mutually rational cooperative behaviour.

The concept of an iterated game, however, opens up a greater range of strategy options to the players involved, since various combinations of C and D may be employed over time. In order to demonstrate this, it will be necessary to define the terms and assumptions at play in the iterated prisoners' dilemma analysis. The bulk of the following material refers to a 2-player game theoretical set-up, but I shall later provide comment about the extrapolation to *n*-player prisoners' dilemma scenarios.

Let us, as Taylor and others do, use the term *supergame* to denote a series of games in which the strategy choices and payoffs available to players remain fixed.[54] A prisoners' dilemma supergame, then, is an iterated interaction in which players may either Cooperate or Defect, and in which their preferences remain those of the one-shot game. We shall refer to each game in the series that makes up the supergame as a *constituent* game, and assume that they are played at "regular discrete intervals of time."[55]

It is an assumption of the supergame analysis that each player has knowledge (or memory) of the players with whom he has previously interacted, and of their strategy choices in such constituent games. An assumption of this type seems necessary for the supergame to move beyond the argument provided by the one-shot prisoners' dilemma; to assume that players have no knowledge of past iterations would be akin to positing that all interactions occur between strangers. The strength of such an assumption, however, is a contentious issue, and one I shall return to presently.

It should by now be clear how the prisoners' dilemma supergame allows for far more (and more complex) strategic choices than the one-shot model. The knowledge assumption above means that players have conditional strategies available to them; strategies which depend on how one's fellow player behaved

[54] Of course, as Taylor notes, we may consider a supergame in which payoffs are dynamic over time to be explanatorily preferable to one in which they remain fixed. Providing such a model is beyond my remit here. See Taylor, *The Possibility of Cooperation* (Cambridge University Press, 1987) p107.
[55] Ibid, p61.

in previous iterations. This proliferation of available strategies, as we shall see, is significant for the arguments analysed in this chapter. Knowing that one's strategy choice in any constituent game may affect how others interact with you in future games may lead one to behave differently; the factor of reputation, one might think, may lead to fewer attempts at exploitation in prisoners' dilemma scenarios. As Axelrod points out;

> What makes it possible for cooperation to emerge is the fact that the players might meet again… The future can therefore cast a shadow back upon the present and thereby affect the current strategic situation.[56]

Note, however, that not all of the new strategy choices available rely on the possibility of basing one's decisions on such knowledge; one may, for instance, choose to Defect unconditionally, or to employ a strategy which alternates between C and D no matter what the other player does.

The last contract problem and indefinite iteration

In addition to the assumptions stated above, we may also posit that players know (or at least believe), prior to the start of the first constituent game, that the interaction will be repeated into the future. Without such knowledge, we would again be left with a framework that hardly differs from the one-shot analysis; players will treat every game as a stand-alone interaction, and mutual defection will occur.

This point, however, forms the basis of an important initial objection to a primitive statement of the kind of iterative story under scrutiny here. Consider the following example. The hedge between Anne's house and Bill's house requires cutting, and the job is one that can be completed by one or two people. Anne prefers that Bill do all the hard work while she reads her book, and Bill prefers the opposite. Each of them, however, would prefer to contribute half of the workload (to ensure that the hedge does get cut) than to leave the hedge unkempt. However, each would rather leave the hedge as it is than get exploited

[56] Axelrod, *The Evolution of Cooperation - Revised Edition* (Basic Books, 2006), p10.

by the other. These preferences, then, are clearly those of a prisoners' dilemma, and may be portrayed in the following (ordinal) preference matrix;

		Bill	
Anne		Cooperate	Defect
	Cooperate	2,2	4,1
	Defect	1,4	3,3

On the one-shot analysis, Defect is the dominant strategy for each player, and so mutual defection is the outcome; the hedge is left to grow and each player receives their third-best payoff.

What, though, if we introduce a basic notion of iteration? Let us imagine that the hedge in question requires cutting just once per year, and that Anne and Bill will face this same decision scenario on more than one occasion. Each of them, as noted above, is assumed to have knowledge of previous constituent games and the other player's strategy on those occasions. Let us posit, for the sake of my argument here, that Anne has arranged to move house in the sixth year after the start of this hedge interaction, and that both parties are certain that this move will occur. The supergame, then, will consist of exactly five constituent games.

Mutual defection across this five-game series will give each player their third choice preference five times. It is clear that both players could do far better in the supergame, either through mutual cooperation throughout or by other combinations of strategies which, although Pareto-inferior to constant cooperation, may still be Pareto-preferred to defection throughout. Whether such cooperation may emerge will be properly discussed later in this chapter; pertinent here is simply to observe that, in the example described, even this possibility seems to be pre-empted.

Anne, knowing that the fifth constituent game will be the last, knows that her decision in that game will have no impact on Bill's future behaviour, or at least that if it does, she will not be around to suffer disutility from it. On the fifth game, as Axelrod notes, "there is no future to influence."[57] Such thinking, however, will not lead to Anne free-riding, since Bill applies the same reasoning

[57] Ibid, p10.

and defects too; mutual defection occurs in the fifth game. This kind of thought-process does not stop here; anticipating unconditional noncooperation in the fifth game, both players see that their conduct in the fourth game has no effect on future interactions, and thus defect in that iteration too. The reasoning applies all the way back to the first game, and precludes the very possibility that cooperation develop between the two parties.

This issue, which we might label the "last contract problem," can, according to Luce and Raiffa, apply formally to all supergames of determinate length;

> The argument generalises: Suppose [the players] know that [the game] is to be played exactly 100 times. Things are clear on the last trial, the [D,D] response is assured; hence the penultimate trial, the 99th, is now in strategic reality the last, so it also evokes [D,D]; hence the 98th is in strategic reality the last…etc. This argument leads to [D,D] on all 100 trials.[58]

If players know, then, that a supergame is *n* games long, the kind of reasoning that leads to mutual defection in the n^{th} game works back through the (n-1)th, the (n-2)th, and so on, such that cooperation does not occur at all.

To avoid this ruling out of cooperative behaviour, the iterative accounts in question are commonly based on supergames that are of *indefinite* length; players simply do not know at the outset how long their interactions with others will last. This assumption, I venture, is appropriate to the modelling of many real life interactions, since in most instances we are uncertain of how far into the future our exchanges with others will last. Admittedly, there may be cases in which the number of constituent games is fixed (such as in the Anne and Bill example), but it seems arbitrary to rule out the possibility of cooperation in all supergames on this basis.

There may remain, I think, a slight issue with the notion of an indefinitely long supergame; I shall return to this point in the latter part of this chapter. For

[58] Luce & Raiffa, *Games and Decisions* (New York: Wiley, 1957), p98-99.

now though, allow us to set aside the last contract problem, and proceed with our exposition of the iterative solution to the prisoners' dilemma.

Discounting payoffs over time

A further key assumption of the accounts of both Axelrod and Taylor is that agents value (otherwise equivalent) payoffs differently depending on when they will be obtained. As Taylor puts it, it "is reasonable to assume that the present worth to a player of a future payoff is less the more distant in time the payoff is to be made."[59] Axelrod states two main reasons for thinking that this might be the case;

> The first is that players tend to value payoffs less as the time of their obtainment recedes into the future. The second is that there is always some chance that the players will not meet again. An ongoing relationship may end when one or the other player moves away, changes jobs, dies, or goes bankrupt.[60]

In order to calculate the potential utility that a supergame strategy offers, then, it seems inappropriate to simply total the cardinal payoffs one expects to receive throughout the constituent games; those that one receives later in the series will simply not be as important to one at the time of making (or, indeed, revising) one's strategy choice. To assume otherwise, claims Taylor is "quite implausible."[61]

Axelrod and Taylor take these observations into account by applying the notion of *discount* to the prisoners' dilemma supergame. The value of future payoffs is assumed to decrease exponentially; one's next payoff is always worth some (unchanging) fraction of the previous one. Let Taylor's term *discount factor* (henceforth DF) denote that fraction; a player for whom each subsequent payoff is worth 90% of the previous one has a discount factor of 0.9.[62] An

[59] Taylor, *The Possibility of Cooperation* (Cambridge University Press, 1987) p61.
[60] Axelrod, *The Evolution of Cooperation - Revised Edition* (Basic Books, 2006), p12.
[61] Taylor, *The Possibility of Cooperation* (Cambridge University Press, 1987) p62.
[62] Ibid, p61.

agent's *discount rate* (henceforth DR) is given by 1-DF, and it is assumed that DF is a number higher than zero but lower than one.

Before proceeding to a demonstration of how discounted payoff calculation functions, it is worth noting one further point. The discounting of future payoffs allows us to calculate cumulative supergame payoffs at the outset without the difficulties that may be thought to arise in the analysis of indeterminately lengthy supergames. Initially, it may appear to allow us to avoid the disturbing prospect that a player be assigned infinite expected utility. If a player's payoffs stretch into the future without any specified cut-off point, one might posit that his constituent payoffs are infinitely repeated. This is problematic, though, because it renders mute the notion of rational strategy in game theory; all cumulative payoffs will be infinite, since any constituent payoff (be it that of mutual cooperation, mutual defection, unilateral cooperation, or unilateral defection) will be multiplied by infinity.

It is at least unclear whether such a worry is pertinent. Recall that the supergame under scrutiny is one of indeterminate, and not infinite, length; games are assumed to end at some stage, players are simply unaware of when. Given that this is the case, it seems doubtful whether one should be concerned about the assignment of infinite expected utility. More relevant is surely the fact that, in indefinitely iterated prisoners' dilemmas, supergame payoffs will be impossible to calculate; since players don't know how many iterations will take place, they simply cannot arrive at an expected payoff figure.

Discounting solves this problem. Due to being exponentially discounted, utility payoffs will approach a certain number, which is taken to provide a satisfactory metric for choosing between potential supergame strategies. Of course, a player's actual cumulative (non-discounted) payoff may be less or more than this figure, depending on how many constituent games there turn out to be. What is relevant is that discounting provides a way to compare and choose between supergame strategies at the outset, in such a way which is sensitive to the fact that the value of payoffs recedes as they recede into the future.

The discounted supergame payoffs for certain strategies, are worked out in the following manner. Assume that the following (arbitrarily chosen) cardinal prisoners' dilemma payoffs are the same for both players; mutual cooperation = 3, mutual defection = 1, unilateral defection = 5, unilateral cooperation = 0.

Assume that each player's DF is 0.8. Mutual cooperation throughout the supergame, then, would lead to a supergame payoff for each player which is the sum of the following infinite series;

$3 + (0.8 \times 3) + (0.8^2 \times 3) + (0.8^3 \times 3) + \ldots$

Or;

$3\,(0.8 + 0.8^2 + 0.8^3 + \ldots)$

The sum of such a series is given by (original payoff) / DR, which in this case is 3/0.2, or 15. Mutual defection, on the other hand, would give each player a supergame payoff of 1/0.2, or 5. With this assumption in place, I may now turn to the account offered by Robert Axelrod.

Axelrod's computer tournaments

Robert Axelrod, in *The Evolution of Cooperation*, provides a novel way of approaching the problem posed by the prisoners' dilemma. Axelrod argues that cooperation may emerge within the context of the indefinitely repeated game, and seeks to demonstrate as such through an analysis of the results of two computer tournaments.

Axelrod begins by noting that the vast amount of literature on the prisoners' dilemma fails to reveal a great deal about "how to play the game well" in its iterated form.[63] Experimental approaches to the repeated prisoners' dilemma, Axelrod posits, are inadequate, due to the players' lack of familiarity with the supergame. "Analysing the choices made by players who are seeing the formal game for the first time," he points out, is likely to restrict "their appreciation of the strategic subtleties" available to them.[64] Learning how to choose effectively (and thus, given the game theoretical framework, rationally) in prisoners' dilemma supergames, thinks Axelrod, requires a new method, one

[63] Axelrod, *The Evolution of Cooperation - Revised Edition* (Basic Books, 2006), p29.
[64] Ibid, p29.

which draws on "people who have a rich understanding of the strategic possibilities inherent in a non-zero-sum setting."[65]

That agents know about the strategic possibilities available to them is usually, of course, simply assumed in normal game theoretical analysis; the additional supergame assumption that players have knowledge of the history of their repeated interaction is intended as another adjunct to this presupposition. Axelrod makes this point, I think, because his account falls between a formal argument and the kind of experimental approach he mentioned above. The "people" in question are neither the subjects of empirical prisoners' dilemma experiments, nor the ideally rational players of formal analyses; rather they are theorists whom Axelrod invited to take part in his computer tournaments.

A range of academics familiar with the prisoners' dilemma were invited to submit computer programmes that embodied supergame strategies, such that all entrants knew that other strategy rules would be created by suitably informed parties. Programmes were then tested pairwise in a round-robin tournament; against all other entrants, against their own twin, and against a rule which chose at random between C and D on each constituent game. In the first tournament, each supergame was set at 200 iterations in length, and repeated five times to "get a more stable estimate of the scores for each pair of players."[66] In the second tournament, the end points of games were determined probabilistically. Fourteen programmes were entered into the first tournament, sixty-two in the second. In each, payoffs were assumed to be those I used in the previous section; 3 for mutual cooperation, 1 for mutual defection, 5 for unilateral defection, and 0 for unilateral cooperation.

A full exposition and interpretation of the results of Axelrod's study is beyond my remit here, but we may note the main conclusions that he draws from it. The most successful strategies (in terms of average supergame payoff), Axelrod points out, were those which generally exhibited four principle features. Firstly, they were *nice*; they were programmes which didn't make the first defection in the supergame series. This allows interaction with fellow 'nice' strategies that usually guarantees mutual cooperation throughout the game. Such niceness, however, can leave a strategy open to exploitation. A second feature –

[65] Ibid, p30.
[66] Ibid, p31.

that of being *retaliatory* – "discourages the other side" from persistent defection; a retaliatory programme responds to a defection with defection of its own for a certain number of constituent games.[67] The third feature is that of *forgivingness*; a propensity to give one's fellow player another chance at cooperation after one 'punishes' them; this allows cooperation to restart even after unwarranted exploitation. Finally, the successful strategies were relatively clear.

The programme which best embodied these features was TIT FOR TAT (henceforth TFT), a conditional strategy which cooperates in the first constituent game, and thereafter does whatever the other player did on his previous move. Against an unconditional defector, for example, TFT would choose C in the first iteration, and D in all subsequent rounds. This strategy 'won' both tournaments; outscoring all other entrants in terms of average payoff over the series of supergames. In the first tournament, for instance, in which programmes could score between 0 and 1000 (due to there being 200 constituent games), TFT gained an average score of 504.5.

At this stage, it may appear that the iterative approach to the prisoners' dilemma fails to move beyond what was already part of the definition of the one-shot analysis; namely that mutually cooperative behaviour provides greater utility than does mutual defection. In the one-shot case, mutual defection arises not because parties prefer it to mutual cooperation, but rather as a consequence of individual rational choice. If the supergame analysis is to convince us that cooperation is possible in prisoners' dilemma scenarios, it must say more about how and when strategies that lead to such outcomes can be rational for agents in the supergame. Such an argument, I think, is best provided by Michael Taylor, to whose account I now turn.

Taylor and supergame equilibria

In *The Possibility of Cooperation*, Michael Taylor works out the conditions under which the combinations of certain types of strategy (such as Tit-for-Tat, unconditional defection, and unconditional cooperation) form Nash equilibria in the supergame. A Nash equilibrium, recall, is a "strategy vector such

[67] Ibid, p54.

that no player can obtain a larger payoff using a different strategy while the other players' strategies remain the same."[68] An obvious example of such an equilibrium in the prisoners' dilemma supergame occurs when both players employ a strategy of unconditional defection throughout the iterations (a strategy Taylor refers to as D^{∞}); a unilateral switch to any other strategy "must either result in D being played on every move (in which case switching it unilaterally from D^{∞} yields the same payoff) or in C being played in one or more constituent games."[69] This move doesn't effect the other player, who defects unconditionally, so in these games when he plays C, the player gets a worse payoff than if he had stuck with D^{∞}. (D^{∞}, D^{∞}), then, is always a Nash equilibrium in the supergame. Unconditional cooperation, meanwhile, is never an equilibrium point, since it will always pay one of the players to switch to a strategy in which they defect in at least one constituent game.

What, though, about conditional strategies? Against a player using D^{∞}, one would do better using D^{∞} oneself than by using TFT; but if both players employ TFT, mutual cooperation occurs throughout the supergame. "Can this," asks Taylor, "be sustained; that is, is [(TFT, TFT)] an equilibrium?"[70]

The answer, it emerges, is that it can be, under certain conditions. Let us, for the sake of clarity, use the prisoners' dilemma matrix employed by Taylor;

	C	D
C	*x, x*	*z, y*
D	*y, z*	*w, w*

(Where $y > x > w > z$.)

We may first consider whether it pays to unilaterally defect from TFT to D^{∞} when faced with another TFT player. If the first player did so, then the game outcomes would change from CC, CC, CC, CC, CC… to DC, DD, DD, DD, DD… Whether such a change in strategy yields a gain for the first player depends on his *discount rate*, because he stands to gain utility in the first iteration, but lose in every subsequent constituent game compared to how he

[68] Taylor, *The Possibility of Cooperation* (Cambridge University Press, 1987) p63.
[69] Ibid, p65.
[70] Ibid, p66.

would have done with TFT. The strategy change yields a net gain only if "he valued later payoffs so much less than earlier ones that the gain in the first period outweighed the losses in all later periods."[71] Whether or not it is rational to maintain TFT over D^{∞}, then, depends on a player's discount factor, relative to the size of the cardinal payoffs on offer. Taylor works out that such a switch to D^{∞} doesn't pay when the following condition is satisfied;

$$DF \geq y - x \ / y - w$$

or, when put in terms of discount *rate*;

$$DR \leq x - w / y - w$$

Let us posit, for the instant, that $y = 5$, $x = 3$, $w = 1$, $z = 0$. Switching from TFT to D^{∞} (when one's fellow player is employing TFT) yields a benefit only if the player's discount rate is strictly more than (3 – 1) / (5 – 1), or 0.5; i.e. when each future payoff is worth less than half of the current one to that player. When the player's discount rate is 0.5 or lower, he will do better by remaining with the TFT strategy in this case.

So under certain conditions, the strategy pairing (TFT, TFT) is "robust against unilateral defections to D^{∞}."[72] This does not, of course, guarantee that (TFT, TFT) is an equilibrium point when such conditions are met, since it might still benefit one of the players to defect to a strategy other than D^{∞}. Taylor considers the possibility that a player defect to a strategy which is identical to TFT, except that it involves defection rather than cooperation in the first constituent game. Let us label this strategy TFT2. If player one defects to TFT2, the sequence of game outcomes would change from CC, CC, CC, CC... to DC, CD, DC, CD...

Defection to TFT2 (and the pattern of alternating unilateral cooperation which it creates) fails to benefit the player if and only if;

$$DF \geq y - x / x - z$$

[71] Ibid, p66.
[72] Ibid, p67.

So, as Taylor notes, "a *necessary* condition for [(TFT, TFT)] to be an equilibrium is that both players' discount factors...must be at least as big as the larger" of $[(y-x)\ /(y-w)]$ and $[(y-x)/(x-z)]$.[73] It turns out that this is also a sufficient condition; since no other candidate strategy can possibly do better against a TFT player than those considered so far (TFT2, D^{∞}, and TFT itself). Recall that a Nash equilibrium point is such that each player's strategy is the "best response" to the strategy of the other. Taylor considers such possible responses to TFT.

Let us momentarily suppose that the best response to TFT is a strategy that plays C on the first move. If this is the case, the TFT player will employ C on his second move, and the best response must be to play C again. This is because;

(1) for each player the game ahead is the same at any point in time (since both the discount rates and the constituent game payoffs were assumed to remain constant)…
(2) a player using strategy [TFT] is responsive to the other player's choices in the preceding game only...It follows that, if the best response to [TFT] begins with C, it must play C in every constituent game. [74]

This being the case, there is no strategy that can improve on TFT itself as a response to a TFT-playing opponent.

Now let us suppose that the best response to TFT is a strategy that plays D on the first move (and hence that whenever the TFT player Cooperates, D is the best response). If this is the case, the TFT player will employ D on his second move. If the best response to TFT involves playing C when the TFT player plays D, then outcome will be an alternating pattern of CD, DC, CD, DC...etc., which is that generated by defection to TFT2. If the best response involves playing D when TFT plays D, then mutual defection is generated throughout the supergame; the series of outcomes that defecting to D^{∞} would ensure.

[73] Ibid, p68.
[74] Ibid, p68.

The above observations show that no strategy can be a better response to TFT than the best of TFT, TFT2, and D^{∞}. The necessary condition above – that both players' discount factors must be at least as big as the greater of $[(y - x) / (y - w)]$ and $[(y - x) / (x - z)]$ – is therefore also a sufficient one for (TFT, TFT) to be a Nash equilibrium point in the supergame. If defection from TFT to neither TFT2 nor D^{∞} improves one's supergame payoff when faced with a TFT player, then no defection will offer an improvement.

Equilibrium and outcome

By Taylor's own admission, his analysis of possible equilibrium points within the iterated prisoners' dilemma is an incomplete one, since he does not attempt to consider every possible set of strategies. He does provide a brief discussion of the possibility of other cooperative equilibrium points within the supergame, but for our purposes here it shall suffice to limit our attention to the three strategies outlined above; TFT, TFT2, and D^{∞}. The arguments considered thus far show that the strategy pair (TFT, TFT) is, under certain conditions relating to cardinal payoffs and discount factors, is a Nash equilibrium within the supergame.

What though, of the possibility that there be more than one equilibrium point in a given supergame? Taylor's analysis also shows that all of (TFT2, TFT2), (TFT, TFT2), (TFT2, TFT) may also be equilibria, under certain conditions; none of which necessarily negate the equilibrium conditions of (TFT, TFT). The pairs (TFT, TFT2) and (TFT2, TFT) create patterns of alternating unilateral cooperation, whilst the pair (TFT2, TFT2) leads to mutual defection throughout. Why think, then, that the cooperative pattern of behaviour would emerge as the actual outcome in such cases?

The fact that a Nash equilibrium point may not emerge as the actual outcome of a game seems to be obscured when we focus on cases with a single equilibrium. As Taylor notes;

> …an equilibrium is such that, if every player expects it to be the outcome, and therefore expects all other players to choose the strategies appropriate

> to this equilibrium, he has no incentive to choose other than his equilibrium strategy; so the equilibrium will be the outcome. [Therefore] if a game has only one equilibrium… it would in fact be the outcome.[75]

If the conditions for two (or more) equilibria are simultaneously met, however, then "the fact that a certain strategy pair is an equilibrium is not in itself a sufficient reason for any player to expect it to be the outcome."[76] Given this, one may doubt whether a given cooperative equilibrium would actually be the outcome among rational agents. Is such a worry justified?

Recall that the pair (D^{∞}, D^{∞}) is also *always* an equilibrium in the supergame (since no player can respond better to a D^{∞} player than by employing D^{∞} themselves). Given this fact, we can locate the potential worry even without searching for cardinal payoff and discount factor values that satisfy the equilibrium conditions for more than one of the above strategy pairs. If (D^{∞}, D^{∞}) is always a Nash equilibrium, when and why should we think that the cooperative equilibrium (TFT, TFT) will be the outcome, even when its equilibrium conditions are met?

Conveniently, when (TFT, TFT) is the only equilibrium point aside from (D^{∞}, D^{∞}), the question is already settled by the structure of the prisoners' dilemma; all parties strictly prefer mutual cooperation to mutual defection, and so the cooperative pair (TFT, TFT) will be the outcome. When the conditions for such a strategy pairing to be an equilibrium are met, then, cooperative behaviour occurs instead of the noncooperative sequence.

What, though, of the possibility that the conditions for the other possible equilibria are also met? Firstly, consider the case in which both (TFT, TFT) and (TFT2, TFT2) are equilibrium points. The latter leads to mutual defection throughout the supergame, and thus creates the same pattern as (D^{∞}, D^{∞}). As such, all parties prefer (TFT, TFT) to (TFT2, TFT2), and thus the former will be the outcome.

More problematic is the case in which the equilibrium conditions of (TFT, TFT) and both (TFT, TFT2) and (TFT2, TFT) are met. The latter two pairs, recall, create patterns of alternating unilateral cooperation. Such patterns

[75] Ibid, p78.
[76] Ibid, p78-79.

can, depending on cardinal payoffs and discount rates, allow parties to do better than they would if mutual defection occurred throughout the supergame; yet they can hardly be regarded as *cooperative* outcomes in any relevant sense. An interaction in which parties take turns to exploit one another looks, prima facie at least, an inappropriate basis for the claim that coercive enforcement is unnecessary. Firstly, it seems hard to imagine how this kind of interaction would get started and become established among individually rational agents. Secondly, it seems feasible that such interactions would multiply the chances of free-riding and exploitation, in a manner which may be thought to warrant the intervention of something like a state.[77]

The likelihood that (TFT, TFT), and the pair (TFT, TFT2) and (TFT2, TFT) are all equilibrium points within the supergame, Taylor notes, is very remote; for this to be so, each player's discount factor must be exactly $(y - x) / (x - z)$.[78] This occurrence is so unlikely, Taylor suggests, that we may simple "ignore this possibility."[79] What, though, if the condition was met? Taylor admits at this point that the outcome would be inconclusive, since the TFT2 player would be indifferent between sticking to TFT2 and employing TFT. This possibility, though, only harms Taylor's argument in a minor fashion; the fact that, given certain (very) specific conditions on payoffs and discount rates are met, a cooperative pattern may not the outcome even if it is an equilibrium, is something Taylor can simply accept. Since such scenarios will rarely arise (and when they do, cooperation *may* still occur), his main point – that cooperative outcomes may arise from rational choice in the supergame – still stands.

The general conclusion Taylor draws from his analysis, then, is that "if each player's discount rate is sufficiently low, the outcome will be mutual cooperation throughout the supergame."[80] When certain conditions concerning cardinal payoffs and discount rates are met, cooperative behaviour can be

[77] These points are stated merely to hint at potential problems that I see with regarding (TFT2, TFT) and (TFT, TFT2) as cooperative equilibria. It may be argued that the two strategy outcomes in question do constitute cooperation of a type stable enough to undermine the game theoretical justification of the state. If this were shown to be so, my position would be strengthened; the possibility of a larger set of cooperative equilibria within the supergame would further undermine the notion that noncooperation is the only possible outcome.
[78] Ibid, p80.
[79] Ibid, p80.
[80] Ibid, p81.

rationally required, and emerge as the actual outcome, in prisoners' dilemma supergames.

The *n*-player supergame

Taylor also extends his analysis to the *n*-player prisoners' dilemma supergame; an iterated game "whose constituent games are Prisoners' Dilemmas with any finite number (*n*) of players."[81] Quite clearly, such a model is well placed to represent many real world prisoners' dilemma-type interactions, which commonly occur among groups larger than two people. As such, this analysis is a clear departure from the work of Axelrod, whose account was based on the aggregated results of pairwise interactions.

An *n*-player game is, in many ways, an extrapolation of the two-player format. As such, the following conditions must be met for it to be a prisoners' dilemma;

(1) The defect strategy dominates the cooperate strategy for every player.
(2) All players prefer complete mutual cooperation (among *all* parties) to complete mutual defection.

In the *n*-player game, a given player's payoff is a function of his strategy, and of the number of other players who choose C in that constituent game.

In such games, as in the two-player game, a situation in which all parties are unconditional defectors is always an equilibrium; no agent can improve their (immediate or cumulative) payoff by defecting to a strategy that involves one or more cooperative moves. Similarly, unconditional cooperation among all parties is never an equilibrium; one would always do better by switching to a strategy which involves defection.

Cooperation, claims Taylor, can occur no matter how many players there are in a supergame. It may do so in a similar way to its occurrence in the two-player game; through the employment of strategies that are conditional on the actions of the agents with whom one interacts. For cooperation to emerge, it must be "conditional on that of *all* the other players," and discount rates must not be

[81] Ibid, p82.

"greater than a certain function of the constituent game payoffs."[82] Further, even when some players insist on using D^{∞}, cooperation can still be rational for the remaining players, if their cooperation is conditional on the cooperation of all other cooperators, and discount rates are sufficiently small.[83]

"Cooperation," Taylor admits, "amongst a large number of players is 'less likely' to occur than cooperation amongst a small number."[84] This is firstly because more conditions need to be met when there are more players in the supergame, and secondly because, in more populated games, the kind of "monitoring" of others' strategies that is necessary for conditional cooperation becomes more difficult. This admission is a telling one, and will be revisited presently.

Taylor's conclusion to his n-player analysis, then, is similar (if rather more complex) than the lesson we may draw from the two-player game. Again, when certain conditions regarding payoffs, discount factors, and game size are met, conditionally cooperative strategies can arise as Nash equilibria in the prisoners' dilemma supergame. As such, they are likely to be rationally selected by agents seeking to maximise their supergame payoff. The upshot of this is that cooperation is, at the very least, more likely than it is taken to be in the game theoretical justification of the state.

Objections to the iterative project

(i) Theoretical possibility and real public goods scenarios

It might be argued that Taylor's conclusions, although interesting, fail to convince us that cooperation will ever occur in the kinds of scenarios we have in mind. It tells us that cooperation is possible if and when certain conditions – regarding cardinal payoffs and discount rates – are met. Yet it is largely silent on whether or not they *are* met in the kinds of public goods scenarios with which he is primarily concerned. Howard Harriott makes a similar point, noting that

[82] Ibid, p104.
[83] Such 'cooperative subgroup' equilibria, however, will not necessarily emerge as outcomes, since "there are many such equilibria, and each player prefers an outcome in which he is an unconditional defector to one in which he is a cooperator.
[84] Ibid, p105.

"because of the abundance of Nash equilibrium points, we do not have enough of a rationale for supposing that the cooperative solution will be uniquely chosen."[85]

Has Taylor not done enough, though, by showing the mere possibility of cooperation in such contexts? Wasn't his aim simply to demonstrate that the common analysis of the prisoners' dilemma unfairly rules out the very possibility of mutually beneficial interaction? His argument, recall, is set up against those who argue that the provision of public goods without external enforcement is impossible. On a sympathetic reading, Taylor need not present a positive argument to the effect that payoffs and discount factors in certain empirical situations are such that the equilibrium conditions for cooperation are met; the weaker conclusion that they *may be* met is sufficient for his purposes. We may doubt, I think, whether the further task – that of showing that cooperative equilibrium conditions are met in some of, or most of, or all empirical prisoners' dilemma scenarios – is even one that could be tackled. At the very least, it would seem to be beyond the remit of Taylor's theoretical project.

Taylor's opponent, however, may still have a gripe, even if he grants both the possibility of cooperation, and the difficulty of identifying all those cases in which the equilibrium conditions for cooperation are met. It may be argued that the equilibrium conditions in question will be met so infrequently as to render the provision of public goods extremely rare. Say, for instance, that mutual cooperation would occur in only 1% of repeated games in which (ordinal) preferences were those of the prisoners' dilemma. The state, if this were the case, may still have some role; to produce such goods in the remaining scenarios. So while it may demand too much of Taylor to require a comment on exactly which empirical scenarios meet the criteria, it might be prudent for him to (even roughly) suggest the level of frequency with which we might expect the cooperative equilibrium conditions to be met in public goods interactions. Given Taylor's general conclusion ("if each player's discount rate is sufficiently low, the outcome will be mutual cooperation throughout the supergame"[86]), it may indeed be reasonable to expect some remark about whether or not the discount

[85] Harriott, 'Games, Anarchy, and the Nonnecessity of the State,' in Sanders & Narveson ed., *For and Against the State* (Rowman and Littlefield, 1996) p129.
[86] Ibid, p81.

rates of actual agents are, in general, such that cooperation could emerge on a significant scale.

(ii) Community and reciprocity

The conclusion to Taylor's n-player analysis may be thought to raise another issue. He holds that the monitoring of other individuals' strategies necessary for cooperation is more likely in a small group, especially one "with an unchanging or very slowly changing membership, or in a community.[87]" Firstly, it is worth noting that those who attempt to justify the coercive actions of the state on the basis of the game theoretical arguments under examination here may well accept such a claim; they may merely assert its irrelevance, given that the groups for which states provide public goods are generally large, and with rapidly changing memberships. Someone sympathetic to Taylor's approach could reply that the mere empirical fact that public goods have been provided for large groups of people (by states) does not ground the further normative claim that this ought to be so.

Let us leave this line of thought aside, however, and focus on Taylor's notion of community. Taylor takes community to include a certain "quality of relations" between members, characterised by (1) shared beliefs and values, (2) direct and many-sided relations, and (3) the practice of generalised reciprocity.[88] These features, at first glance, seem to be suggestive of some quasi-cooperative interactive situation between community members. If cooperation is only feasible when such factors are in place, we may wonder whether Taylor's cooperative solution to the prisoners' dilemma isn't parasitic on some pre-existing cooperation in a given community. It may even appear that Taylor is begging the question in this manner.

We may note that this kind of worry is particularly pressing if one conceives of the problem of cooperation as fundamentally a state of nature problem. In man's pre-political state, one could argue, community itself is unlikely to form due to the very inability of individuals to cooperate. It looks

[87] Taylor, *The Possibility of Cooperation* (Cambridge University Press, 1987) p105.
[88] Ibid, p23.

problematic, in this respect, to place community as the primary context in which cooperation can emerge at all.

Taylor can, perhaps, avoid such a criticism by fleshing out the kind of communal relations he has in mind; there may not be a problem if the conditions don't include an already-established system of cooperation in the formal sense. His third feature in particular seems to require such expansion; "generalised reciprocity" could very easily be taken to indicate a situation in which some mutual cooperation is already in place. Taylor needs to convince us that this is not so.

Further inspection of Taylor's position, however, merely seems to exacerbate the problem. In *Community, Anarchy and Liberty*, Taylor states the following;

> I shall use [the term reciprocity] to cover a range of arrangements and relations and exchanges, including mutual aid, some forms of cooperation, and some forms of sharing.[89]

The very notion of community that Taylor is using, then, is (in part) characterised by a reciprocity that may already include the existence of some cooperation. An example he gives of the type of interaction he has in mind – "when the primitive or peasant cultivator gives up time to help others harvest crops quickly in the firm expectation that those he has helped will do the same for him" – seems to be exactly the kind of cooperative interaction that is pertinent to this paper, and to Taylor's iterative account. The suggestion, then, that cooperation is far more possible under conditions in which some cooperation may already be present, seems to be a problematic one, and might be thought to impact on the plausibility of his project.

Taylor could react to this point in two ways. The first would be to withdraw the notion of reciprocity from the notion of community, or at least to play down its importance. The more modest claim – that communities, as relatively small and stable groups of individuals, can facilitate the kind of

[89] Taylor, *Community, Anarchy and Liberty* (Cambridge University Press, 1982), p28.

monitoring of other players and their strategies that conditional cooperation requires – seems to be a plausible one.

The second response would involve pointing to the fact that Taylor doesn't present community as the *only* scenario in which cooperation can emerge. Rather than laying down a necessary condition for cooperation in the prisoners' dilemma, Taylor is merely suggesting certain circumstances that can make it more likely. As such, the argument in question may be viewed as separate from, and inessential to, his main claims about the *possibility* of cooperation; if, I have suggested, the link between community and cooperation is a problematic one, the former may be discarded without harming the latter. The apparent contingency of the link in question would undermine the notion that the problem outlined above harms the plausibility of Taylor's iterative account.

(iii) Indefinite iteration, infinite iteration, and the last contract problem

As a formal feature of the prisoners' dilemma supergame, it seems unclear whether the assumption of indefinite iteration is sufficient to solve the last contract problem. We may note that an assumption of *infinite* iteration certainly solves the issue; there will be no final contract, so the reasoning in question simply doesn't apply. Such a presupposition, however, simply doesn't seem to map onto the real world; our interactions with others are temporally constrained in a number of ways, not least by our mortality. If cooperation were only possible in infinitely iterated games, then it would seem prima facie problematic to infer that real life cooperation in prisoners' dilemma scenarios was possible.

Instead, then, the (more realistic) notion of indefinite iteration is commonly employed; players know that the game will end at some point, but not when. It seems doubtful, however, whether this assumption solves the last contract problem in the formal setting. Players know that the supergame will end at some stage, that *some* iteration will be the last. Why couldn't they reason back from that final game, even if the time of its occurrence remains undefined or unknown to them?

Perhaps there are ways around this restatement of the last contract problem. Even if there were not, however, I think that I would continue to find the notion of indefinite iteration to be more suitable than that of infinite iteration, for the reasons outlined above. If the more realistic notion of indefinite iteration clashes with last contract logic, my intuitive reaction would be to question the plausibility of the latter.

Intuitively, I think, we hold such last contract reasoning to be problematic; it seems almost irrational to jeopardise a potentially lengthy period of mutually beneficial cooperation merely to gain the advantage of exploitation in the final iteration. Empirically, I posit, we rarely (if ever) employ such thinking; indeed, we can cite countless real life examples in which people do manage to achieve mutual cooperation over an interaction of fixed length. We would, I think, be unsurprised to see Anne and Bill sharing their hedge-cutting duties, and would hardly regard such behaviour as irrational. Further, it seems safe to say that cooperation is even more likely to arise when interactions are of indefinite length.

Such last contract reasoning, however, is the embodiment of rationality within the game theoretical framework; if players can do better by defecting, they do so, and if they expect their opponent to do so to, it is rational to move one's defection to an earlier point in time. This impasse between our intuitions about the rationality of last contract reasoning and the game theoretical model, then, is suggestive of a greater incompatibility. My objection to last contract reasoning is based on a conviction that it is irrational; yet this very intuition invokes exactly the kind of rationality that is precluded from the prisoners' dilemma framework. This tension, as I shall argue later in the paper, is indicative of the fact that game theoretical analysis is inadequate as a basis for social theory.

(iv) Demandingness of knowledge

We might argue that the requirement that one have perfect knowledge of previous interactions with all other players seems to be too demanding, particularly in the n-player game. Taylor admits as such, noting that the kind of

monitoring necessary for conditional cooperation becomes increasingly difficult when the supergame is made up of more players. Such an assumption, like many of those at play in the game theoretical framework, seems not to map onto our real life interactions; we are unlikely to ever have anything approaching perfect knowledge of others and their strategies.

Despite our limitations, however, we do manage to cooperate with others. This final observation – that our empirical experience of social interaction seems to be at odds with the kind of idealising assumptions at play within the game theoretical framework – is one that I shall revisit in the next chapter.

Lessons to draw from the iterative story

The arguments of the type offered by Taylor, then, are interesting in two respects. Firstly, as suggested in this chapter, they provide us with reasons for thinking that cooperation is possible within prisoners' dilemma scenarios, in a manner overlooked by those seeking to provide a justification of the existence of a coercive state. By extending the game through time, Taylor goes some way to showing that rational individuals can achieve mutual cooperation. The analysis succeeds in formalising some concern for our reputations (thereby acknowledging the importance of one's actions for one's future prospects), as well as allowing for the fact that future, as well as current, payoffs may feature in our strategic reasoning. If certain conditions regarding payoffs and discount rates are met, Taylor demonstrates, cooperative strategy vectors can be Nash equilibrium points in the prisoners' dilemma supergame.

Secondly, Taylor's arguments contribute to an external critique of the use of game theoretical models for the purpose of justifying the state. In insisting that certain temporal constraints on the prisoners' dilemma must be loosened, and even (as above) by adding to the list of demanding assumptions at play, such accounts cast doubt on whether such a framework, being so at odds with our real social experience, can form the basis of normative judgements about the state.

CHAPTER 4 – GAUTHIER AND CONSTRAINED MAXIMISATION

I turn now to another theory that seems to provide a plausible alternative to the coercive solution to the prisoners' dilemma. David Gauthier, in his much-discussed book *Morals by Agreement*, aims to show that it can be individually rational (even when rationality is still understood in the game theorist's simple maximising sense) to act in ways that social cooperation requires.

Gauthier's overall project is an ambitious one; he attempts to explain how a moral theory may be understood to arise out of, and be grounded in, the rationality of individuals. Given the scope of *Morals by Agreement*, it is beyond my remit to offer a complete elucidation of Gauthier's position. It will be sufficient for the task at hand, I believe, to focus on one of his central theses; that cooperative behaviour, given certain assumptions, may be rational for individuals faced with prisoners' dilemma-like scenarios.

If Gauthier's 'moral solution' to the prisoners' dilemma is successful, it would provide a way for parties to reach mutually beneficial payoffs without third-party enforcement. It would, therefore, debunk the claim that coercive force of some kind is a necessary condition for cooperation between utility-maximising individuals.

I will now move on to an exegetical presentation of the key elements of Gauthier's position, explaining how it may, prima facie, appear to solve our problem. I will proceed, however, to offer a number of objections to Gauthier's approach. In light of these criticisms (some of which are more telling than others), I will suggest that Gauthier's project fails to convince on its own terms. This conclusion, however, does not spell the end of my interest in Gauthier's work. Although his arguments (and, indeed, those of the iteration theorists I analysed in the previous section) may fail to provide an internal solution to the prisoners' dilemma, they may, I think, be reformulated as part of a redundancy charge concerning the use of the game.

Individual and joint strategies

Gauthier's response to the pessimistic employment of the prisoners' dilemma begins with the distinction between individual and joint strategies. An

individual strategy, Gauthier claims, is "a lottery over the possible actions of a single actor."[90] An agent who acts on an individual strategy, simply put, chooses merely between the different courses of action available to him. A joint strategy, on the other hand, is a "lottery over possible outcomes."[91] The agent who bases his actions on a joint strategy, then, considers the possible outcomes that could result once his choice of action is combined with the choices of other agents with whom he may interact.

Cooperation, it stands to reason, is inherently linked to the notion of joint strategy. Participation in a cooperative activity (Gauthier uses the example of a hunt) may be thought of as the implementation of a single joint strategy by a number of individual actors. The concept may be extended, suggests Gauthier, to include participation in practices like promise-keeping, in which "each person's behaviour is predicated on the conformity of others to the practice."[92]

An individual, of course, cannot ensure that he acts on a joint strategy, for whether he does so depends on how his fellows act. We might, however, say that "an individual bases his actions on a joint strategy in so far as he intentionally chooses what the strategy requires of him."[93] To cooperate, for Gauthier, is simply to employ a joint strategy rather than an individual one.

Gauthier's departure from the kind of rationality that is assumed in the prisoners' dilemma should already be coming into focus at this point. The game theoretical reasoner is one who only employs individual strategies; the question he asks when deliberating over which action to choose is "which of the actions available to me will serve me best?" He who acts on a joint strategy, in contrast, may ask "what outcomes can I and others achieve by coordinating our actions?"

This distinction seems unproblematic, as does the linking of cooperation to joint strategy. He has not yet, though, said anything yet about the rationality of choosing joint over individual strategies; something he must do, given the nature of his project. I shall return to this topic later in this chapter, but allow me now to briefly allude to the issue here. Gauthier admits that the rational maximiser, at this point in his analysis, can reason along the following lines; "it is rational to cooperate only if the utility one expects from acting on the cooperative joint

[90] Gauthier, *Morals by Agreement* (Oxford University Press, 1986), p166.
[91] Ibid, p166.
[92] Ibid, p166.
[93] Ibid, p166.

strategy is at least equal to the utility one would expect were one to act instead on one's best individual strategy."[94] Given the assumption of utility-maximising rationality (one Gauthier must accept if he is to provide an internal solution to the PD), this appears to be a compelling claim; it would surely only be rational to cooperate if doing so afforded one greater utility than would non-cooperation. Gauthier, then, must convince us that cooperation is individually rational, when such rationality is understood in the narrow game theoretical sense.

A problem may arise here. The kind of deliberation needed to ascertain whether cooperating is rational for me in a given case ("will I do better by cooperating?") seems to be of the kind at play in an individual strategy; I am calculating how *I* would fare should I act in different ways. Such thinking, as Gauthier posits, "defeats the end of cooperation, which is in effect to substitute a joint strategy for individual strategies."[95] This may be the case, but to decide whether such a substitution would be to my benefit seems to require the kind of reflection on my potential utility payoffs that is central to an individual strategy. I will return to this thought presently.

Straightforward and constrained maximisers

The elucidation of the difference between individual and joint strategies segues into Gauthier's more fundamental argumentative device; the distinction between straightforward and constrained maximisation. The distinction is first made in the following terms;

> …a *straightforward* maximiser is a person who seeks to maximise his utility given the strategies of those with whom he interacts. A *constrained* maximiser, on the other hand, is a person who seeks in some situations to maximise her utility, given not the strategies, but the utilities of those with whom she interacts. The Foole accepts the rationality of straightforward maximisation. We …accept the rationality of constrained maximisation.[96]

[94] Ibid, p166.
[95] Ibid, p166-167.
[96] Ibid, p167.

Instead of considering merely the strategies of his fellows, like the straightforward maximiser (henceforth SM) does, a constrained maximiser (henceforth CM) considers their potential payoffs; she seeks to maximise her own utility within certain constraints that such consideration implies.

Before we flesh out the crucial CM model, I shall briefly say something about such constraints, which are closely related to Gauthier's considerations about fairness. The notions at play here, particularly that of "minimax relative concession" are somewhat nebulous to those not schooled in formal bargain theory, a number among which I count myself. Sufficient for my purposes, I think, will be Gauthier's own abbreviated summary, which is used in the following precursory characterisation of the CM;

> If he is able to bring about, or may reasonably expect to bring about, an outcome that is both (nearly) fair and (nearly) optimal, then he chooses to do so; only if he may not reasonably expect this does he choose to maximise his own utility.[97]

To say that an outcome is (nearly) fair and (nearly) optimal, Gauthier goes on to clarify, is simply to say that the payoffs it affords to the players are "close to those of the cooperative outcome."[98] Leaving aside the question of optimality (it is part of the structure of the prisoners' dilemma that all players prefer cooperation to universal noncooperation), Gauthier's point appears to something like the following; that the "just man" (Gauthier's term) takes into consideration how fair outcomes would be to each of the parties involved. This is all well and good, but the current debate is one of rationality, and not one of justice. The maximising reasoner, he with whom Gauthier is purportedly engaging, would give such deliberation short shrift; "why should I care about how others do, as long as I get my preferred outcome?" To assume such deliberation, in other words, is to fail to tackle the proponent of the game theoretical argument on his own terms.

This sort of problem will continue to trouble Gauthier, and I shall revisit in section III. For now, though, we may just place the fairness thesis to one side

[97] Ibid, p157.
[98] Ibid, p157.

as an inessential element of Gauthier's project, at least in the context of this paper. The crux of Gauthier's point, as we shall see, is that the CM simply does better than the SM, even when our yardstick is the maximisation of utility. To posit some fairness-related deliberation at the outset seems immediately to distance the argument from the PD framework within which Gauthier wishes to operate.

A CM, claims Gauthier, "gives primary consideration to the prospect of realising the cooperative outcome" when choosing what to do.[99] In terms of the distinction introduced in the previous section, the CM has "a conditional disposition to base her actions on a joint strategy, without considering whether some individual strategy would yield her greater expected utility."[100] Clearly not all such constraint would be rational, so Gauthier goes on to characterise her as;

> (i) Someone who is conditionally disposed to base her actions on a joint strategy or practice should the utility she expects were everyone so to base his action be no less that what she would expect were everyone to employ individual strategies, and approach what she'd expect from the cooperative outcome […].
>
> (ii) Someone who actually acts on this conditional disposition should her expected utility be greater than what she would expect were everyone to employ individual strategies.[101]

Point (ii) here is indicative of Gauthier's attempt to solve the challenge posed by the prisoners' dilemma; given the assumptions at play in the game, an argument for the rationality of cooperation must refer to individual utility maximisation. As Gauthier later says, the CM's disposition to cooperate is "conditional on her expectation that she will benefit in comparison with the utility she could expect were no one to cooperate."[102]

A CM, then, must estimate the likelihood that the other agents involved in a prospective interaction will act cooperatively, and work out the utility she

[99] Ibid, p157.
[100] Ibid, p167.
[101] Ibid, p167.
[102] Ibid, p169.

expects if she cooperates, given the predicted level of cooperation. Only if this is higher than the utility she'd expect from universal noncooperation does her conditional disposition "manifest itself in a decision to base actions on the cooperative joint strategy."[103]

As such, the CM will not play into the hands of people she believes will not cooperate (SMs, for instance). The CM avoids exploitation in such cases by merely employing his best individual strategy; by acting like an SM. A CM makes "reasonably certain" that those with whom she may choose to cooperate are other CMs before actually constraining her direct pursuit of utility.

So Gauthier's CMs can obtain the cooperative benefits that are unavailable to SMs (the former can attain their second-best outcome in the PD, whilst the latter have to make do with their third). CMs do as well as SMs do in interactions with other SMs (by just acting on their best individual strategy), yet do better than SMs in encounters with other CMs (by cooperating). If this were the whole story, then the CM would indeed look more rational than SM in terms of utility maximisation. But Gauthier acknowledges the possibility that a CM fail to recognise an SM (because the SM is posing as a CM, for instance). In such cases, the SM receives his favourite payoff (unilateral defection), whilst the CM is left with the 'sucker' payoff. The question, then, of whether being an SM or a CM is more rational, remains unanswered at this stage.

The rationality of CM as a disposition, then, depends on a CM's ability to identify similarly disposed individuals. If such identification was fool proof – that is, if CMs always correctly recognised other CMs and cooperated with them, and always correctly identified SMs and simply acted on their best individual strategy – then CM would evidently yield greater expected utility than SM, and so would be rationally adopted by agents.

Such identification is complicated, though, by the following line of reasoning; that the most rational thing to do in one's interactions with others is to simply *appear* trustworthy, to *appear to be* to be the sort of agent who is willing to cooperate.[104] As noted above, the possibility that SMs 'trick' CMs is one that could undermine the proposed rationality of the CM disposition. If SMs can gain

[103] Ibid, p169.
[104] Ibid, p173. The reasoning here is similar to that concerning agreements made pre-game, something I discussed in chapter 2.

access to the benefits of cooperative schemes (by pretending to be CMs, for instance), and subsequently exploit those who trust them, it may be more rational to be an SM.

Gauthier defeats this line of reasoning with an idealising assumption; that persons are *transparent*.[105] Accordingly, each actor is "directly aware of the dispositions of his fellows, and so aware whether he is interacting with straightforward or constrained maximisers."[106] This idealising assumption is to stand alongside all of the others at play within the prisoners' dilemma framework, and means that deception on the part of SMs (and misidentification on the part of CMs) is impossible.

The problem, as Gauthier swiftly admits, is that this assumption threatens to rob his account of its force. If CM defeats SM only if individuals are transparent, then "we shall have failed to show that under actual, or realistically possible, conditions," such constraint is rational. To ensure that his account retains its practical relevance, Gauthier plausibly claims that we "must relate our idealising assumptions to the real world."[107]

Transparency, then, is softened to *translucency*. Under the new assumption, an agent's disposition to cooperate or not may be discovered by others, not with absolute certainty, but in a manner that involves more than mere guesswork.[108] Gauthier holds that we are sufficiently translucent for CM to be the more successful disposition. He reaches this conclusion by way of a mathematical formula, which works out the potential payoff gain from being a CM given three probabilities; the likelihood p that CMs will achieve mutual recognition and thus cooperate in a given interaction, the likelihood q that CMs will fail to recognise SMs, but will themselves be recognised, and thus exploited, and the probability r that a randomly selected member of the population in question is a CM.[109]

For CM to be rational, Gauthier works out, probability p must be more than twice the probability q; CMs must successfully identify one another more than twice as much as they get exploited by SMs. This is the case "however great

[105] Ibid, p173.
[106] Ibid, p174.
[107] Ibid, p174.
[108] Ibid, p174.
[109] Ibid, p175.

the probability *r*."[110] To illustrate, Gauthier plugs some numbers into his formula. He calculates that if $p = 2/3$, $q = 1/5$, and $r = 1/2$, CMs may expect to do better than SMs. Given this fact, it would be irrational to continue to be an SM, and so CM would be positively selected by agents.[111]

Even if we accept Gauthier's mathematical rendering of the issues at hand, this seems to me a deeply unsatisfying final step of the argument. I will, however, detail these reservations in part (ii) of the next section. For now, allow me to restate the relevance of Gauthier's proposals here. If his argument works – that is, if CMs really achieve better payoffs than SMs – then cooperation in the PD is rationally achievable even in the absence of a coercive force. This moral solution to the problem looks like one which comes at a far lower cost to us than does the coercive alternative.

Objections to Gauthier's account

I know turn to a series of issues concerning Gauthier's project. The first involves the demandingness of translucency, and is one that I suggest may be solved by supplementing Gauthier's account with elements of the iteration theory discussed in the last chapter. Consideration of a second criticism of translucency will raise a more fundamental concern relating to the final step of Gauthier's argument. I will then discuss a set of objections which focus on the role of dispositions in Gauthier's account. Not all of these criticisms hit their target, but consideration of them will lead us to what I consider the most fundamental problem at hand; that Gauthier will have trouble convincing the individuals in prisoners' dilemma situations of the rationality of CM. I will argue that it is at best unclear whether he can solve the PD from within the game theoretical framework, as he hopes to do. I will conclude that Gauthier's project fails on its own terms.

[110] Ibid, p175.
[111] Ibid, p177.

(i) Translucency and knowledge of dispositions

The requirement of transparency, Gauthier admits, is demanding and unrealistic. Assuming that all of our preferences and dispositions are immediately knowable to others is simply to deny the relevance of the solution to human affairs. Translucency, the assumption with which Gauthier replaces transparency, is undoubtedly more realistic, but may not be able to do the job Gauthier needs it to.

Translucency captures our intuition that individuals may have knowledge (or at least good hunches) about the dispositions of those with whom they interact. As such, the translucency assumption seems to accord with our everyday experiences; we take ourselves to be able to (with at least some accuracy) assess whether those with whom we interact are trustworthy. We do not generally consider the decision whether to cooperate to be a mere shot in the dark.

An issue, though, arises when one focuses the conditional nature of the dispositions. Consider an interaction between two CMs. Player 1's disposition to act according to a joint strategy is conditional upon Player 2 having a similar disposition. But Player 2's disposition is in turn conditional on the disposition of Player 1 (which, as stated, is conditional on the disposition of Player 2, etc.). This conditional interplay between dispositions seems to create (at the very least) a complexity; as Jan Narveson correctly notes; "part of my solving my problem about him is that he has to have solved the same problem about me."[112] Further, the level of convolution is likely only to increase if we consider interactions involving more than two players.

Of course, none of this would be problematic under transparency; which posits that all dispositions (even the most complex) are immediately knowable to others. Under translucency, though, one may wonder whether either party would be sure enough about the dispositions of the other to get cooperation off the ground. If such doubts are salient, then Gauthier would be left with the following dilemma; either posit transparency and risk the practical irrelevance of the

[112] Narveson, 'The Anarchist's Case,' in Sanders/Narveson ed. *For and Against the State* (Rowman and Littlefield, 1996), p208.

account, or posit translucency and risk reducing the likelihood of cooperation between CMs.

Gauthier can, I think, appease his challenger here by invoking the kind of iteration-based story detailed in the previous chapter. Mutual knowledge of dispositions, even conditional ones, is surely more feasible when interactions between the same individuals are repeated over time, and/or occur within the context of some standing relationship. Empirical support for such a point is easy to come by; our everyday friendships provide ample evidence that knowing the dispositions of other actors (even those dispositions which depend on our own) is possible. In fact, I think, Gauthier's framework provides a rather plausible way to characterise the cooperative element of our iterated interactions; Adam disposes himself to cooperate because he knows Bill is disposed to cooperate, and vice versa. By incorporating iteration into Gauthier's account, CMs can more feasibly attain the kind of knowledge that cooperation requires, even under translucency.

A stumbling block, though, is that Gauthier does not want his account to rest on iteration. He expressly states (albeit only very briefly, in a footnote) that two CMs may cooperate "even if neither expects her choice to affect future interaction."[113] Gauthier, claims Anthony de Jasay, "wishes his theory to be perfectly general and not contingent on such particular social facts as iteration."[114] This doesn't seem to provide a convincing reason not to turn to this solution, since Gauthier has elsewhere (when rejecting transparency) demonstrated concern for how his account maps onto real life. While it may be true that iteration-based accounts bring problems of their own, I think that Gauthier ought to accept at this juncture that iteration renders his discussion of dispositions more plausible.

(ii) Sneaky SMs and the charge of arbitrariness

George W. Rainbolt sketches another potential problem with the transition from transparency to translucency. He posits that Gauthier's argument for the rationality of constrained maximisation only holds when a further

[113] Gauthier, *Morals by Agreement* (Oxford University Press, 1986), p170.
[114] De Jasay, 'Self-contradictory contractarianism,' in Sanders/Narveson ed. *For and Against the State* (Rowman and Littlefield, 1996), p153.

assumption is in play; that abilities to detect and conceal one's dispositions are equal.[115] Without such an assumption, Rainbolt claims, there remains the possibility that some agents (what he terms 'Snidelies') be experts at tricking others. This, one may think, would reduce the payoffs available to CMs, since they would be more likely to be exploited by sneaky SMs (and subsequently less likely to trust fellow CMs). In the absence of an equality assumption, then, then translucency will be insufficient to guarantee the rationality of CM over SM; the uncertainty that creeps in when one moves from transparency to translucency will be put to better use by some than by others.

Rainbolt goes on to argue that Gauthier cannot simply add in such an assumption, since it is plainly false; in the real world, "detection and concealment abilities differ dramatically between individuals."[116] Subscription this assumption would harm the force of his argument (much like the assumption of transparency would have).

Gauthier's reply to Rainbolt, I think, would have to run along the following lines; there is no reason to think that the level of translucency exhibited by agents allows for widespread concealment. Gauthier admits that CMs will be exploited on occasion, but the level of identifiability he has in mind is obviously high enough for CMs to be able to avoid the sucker payoff on most occasions. Translucency must be such that inequalities in the abilities in question simply don't make a significant difference to interaction.

This, though, leads us into a new objection; that Gauthier has arbitrarily asked us to accept as realistic a level of translucency at which CM is rational. Recall his example; in a population split evenly between SMs and CMs, CMs could expect to do better if they managed to successfully cooperate with other CMs in 2/3 of their encounters, and avoided exploitation by SMs in 4/5 of their encounters. "These persons," Gauthier concludes, "are sufficiently translucent for them to find morality rational."[117]

---

[115] Rainbolt, 'Gauthier on Cooperating in Prisoners' Dilemmas,' in *Analysis*, Vol. 49, No. 4 (October 1989), p219.
[116] Ibid, p219.
[117] Gauthier, *Morals by Agreement* (Oxford University Press, 1986), p177.

Whilst this may be the case, Gauthier provides little argument to convince us that such figures accurately represent real life interactions.[118] It is clear enough that positing $p = 2/3$, $q = 1/5$ is more realistic than positing $p = 1$, $q = 0$ (which would be the case if transparency, rather than translucency, were true), but this doesn't tell us much. CM, for example, comes out as less rational than SM if you use the figures $p = 2/3$, $q = 2/5$ in Gauthier's formula. The debate, then, would be whether 1/5 or 2/5 would be a more realistic probability to use for q. This task, quite evidently, is a tricky and potentially controversial one. It is, however, one in which Gauthier really needs to engage if he is to truly convince us that CM is more rational than SM.

(iii) Queries over dispositions

Gauthier's use of the notion of disposition is thought by some to generate problems for his account. A preliminary worry may be that dispositions don't seem to be the kinds of things that one may 'choose' in any normal sense; they often seem to be rather fixed elements of human psychology. That I am disposed to eating too much at meal times seems just to be a fact about me, something which influences my choices rather than something that is itself the object of a choice. Thus, Gauthier may be accused of confusing the matter when he speaks of it being rational to an SM to choose to become a CM, for this simply isn't the kind of choice that agents are able to make.[119]

This objection, I think, slightly misses the point. Any idea of a rational 'choice' between the dispositions is merely an argumentative device employed by Gauthier. The point (if Gauthier's argument is successful) is simply that CMs do better than SMs in terms of utility maximisation, and thus that it would be more rational to be a CM. Gauthier need not be regarded as making any claim about the nature of dispositions as applied to real agents.

Holly M. Smith raises the point that Gauthier's model of disposition seems to conflict with one of our intuitions about intention and action. Under the CM disposition, as Gauthier describes it, intentions to act appear to be conclusive

[118] Given his previous claims about how his account relates to the real world, I assume that realism provides the pertinent framework for assessing his claims here. In any case, Gauthier details no other way in which we might judge the suitability of these probabilities.
[119] Ibid p177.

reasons for action; "if I form the intention at t2 of building my lighthouse, then I *will* build it at t3."[120] In other words, there is no possibility that I fail to do what CM requires of me. This strong assumption - one Smith labels the "causal efficacy thesis" – looks to be at odds with our intuitions about such matters.[121] Assuming that intention and action are temporally distinct, we do not generally think it irrational (let alone impossible) to evaluate intentions in light of new facts, and then fail to act. I may intend to drink the glass of (what I at that time think is) water in front of me, but then decide not to do so in the light of new knowledge; suppose that I subsequently become aware that it is a poisonous cocktail.

Gauthier employs such an assumption, we may speculate, in order to make the cooperative disposition binding in some respect; he doesn't want CM to allow for the possibility of an agent intending to cooperate before defecting. In order to rule out such selfish re-evaluation, he has had to also rule out certain practices that we seem to see as rational in a more general sense. In Gauthier's defence, issues such as this one may just be part and parcel of attempting to provide an internal solution to the PD; certain of our intuitions about rationality have no place within its idealised framework. Such a tension, I think, is symptomatic of the fact that the prisoners' dilemma would be best attacked from outside, rather from within, a point I shall return to.

I now turn to an objection raised by Anthony de Jasay, which provides a useful starting point to the discussion of what I consider to be the most fundamental problem with Gauthier's project. De Jasay is puzzled by the concept of a disposition that Gauthier employs in *Morals by Agreement*. He wonders, for instance how mutual dispositions are "able to achieve something that mutual promises cannot."[122] How, de Jasay asks, should the notion of a disposition be understood? It must presumably be "more than a discernable frequency, a statistical probability" to act in a certain way, for such a pattern could arise simply by judging on the merits of each case, meaning that the idea of a disposition, would be doing no real work.

---

[120] Smith, 'Deriving Morality from Rationality,' in Vallentyne ed. *Contractarianism and Rational Choice* (Cambridge University Press, 1991), p235.
[121] Ibid, p235.
[122] De Jasay, 'Self-contradictory contractarianism,' in Sanders & Narveson eds. *For and Against the State* (Rowman and Littlefield, 1996), p154,

De Jasay proposes two interpretations of what a disposition may be.[123] On the first view, a disposition is some element of one's psychological makeup that influences one's preferences, and is factored into one's payoffs. On the second, a disposition amounts to the systematic privileging of one kind of action over and above what the balance of reasons demands. Let us look at each in turn.

According to the first view, a person chooses his actions based on a disposition that "privileges certain ends over others."[124] For example, if an agent rates 'moral' outcomes more highly than his neighbours, then we might say he has a moral disposition. In this sense, one's dispositions are incorporated in the payoffs one expects to attain from different actions.

A CM's disposition, on this interpretation, is such that in interactive situations, he prefers the cooperative outcome to the free-riding one. To say this, however, as de Jasay notes, is just to say that the PD doesn't exist for that agent. To demonstrate, consider the following strategy matrix, which represents a possible interaction between two players with such a disposition (C = cooperate, D = do not cooperate);

		Player 2	
		C	D
Player 1	C	1,1	4,2
	D	2,4	3,3

In such situations, D is no longer the dominant strategy (and DD no longer the sole dominant equilibrium), and cooperation looks at least possible. This much seems uncontroversial. The problem is, however, that this is simply no longer a prisoners' dilemma. As de Jasay suggests, Gauthier's opponent can simply claim that the sentence "noncooperation is the only dominant equilibrium" is simply analytically true of the PD.[125] If this is the case, then Gauthier's express intention – to show how cooperation is rational within the PD – cannot help but fail. If he thinks of dispositions as altering the payoffs available

[123] I think that de Jasay sees these two possible views as exhaustive. Gauthier may wish to dispute this, but de Jasay's point is simply that *Morals by Agreement* does not properly elucidate the notion at play. The alternatives he provides do, at the very least, seem to be two of the most intuitive available.
[124] Ibid, p155.
[125] Ibid, p153.

to players, Gauthier must be seen not as solving the PD, but as replacing it with "a different and less harsh game, albeit one closer to everyday reality."

The second of de Jasay's proposed interpretations is that a disposition consists in the systematic privileging one sort of action over and above what the balance of reason recommends. An agent with a cooperative disposition, then, would simply pursue cooperative strategies in his interactions with others, without regard for whether or not this would be the most rational course of action in each case. This, if we accept the utility-maximising rationality at play within the PD framework (and Gauthier needs to if he's to solve the puzzle from within), simply looks like an irrational mistake; how could it be rational not to assess whether a particular course of would afford me greater utility in each case?

Gauthier provides us with good reason to think that he subscribes to this second notion of disposition. He does so, I think, in the context of an attempt to convince the reader that CM isn't just SM "in its most effective disguise."[126] Gauthier insists that the CM isn't just someone who "serves her overall interest by sacrificing the immediate benefits of ignoring joint strategies…in order to obtain the long term benefits of being trusted by others."[127] Such a person, we presume, would maintain their cooperative actions only for as long as the balance of reasons (over the long term) recommends that they do so. To distinguish his vision of cooperation from this type of reasoning, Gauthier describes the CM as having "a conditional disposition to base her actions on a joint strategy *without considering whether some individual strategy would yield her greater expected utility*."[128] To cooperate without first checking whether cooperation affords one greater utility than non-cooperation, though, as stated, would surely just be ruled out as irrational by the reasoners in the prisoners' dilemma.

Whichever of these notions of disposition Gauthier wishes to employ, then, le is faced with a problem. If he goes down the first path, then it appears that rather than solving the prisoners' dilemma, he merely replaces it with a different game from which cooperation more readily stems. If he endorses the

[126] Ibid, p169.
[127] Ibid, p169.
[128] Gauthier, *Morals by Agreement* (Oxford University Press, 1986), p167. Emphasis added.

second interpretation, the CM seems irrational when judged in terms of the assumptions at play in the PD.

In *Morals by Agreement*, then, Gauthier repeatedly tries to overcome the challenge posed by the game theoretical reasoner. He attempts to do so on his opponent's terms; by providing a cooperative solution internal to the prisoners' dilemma. Despite his efforts, though, and the ingenuity of the CM model, he never truly overcomes a nagging worry; that to constrain oneself, to dispose oneself to act on joint strategies even when one may do better by defecting, looks irrational on the narrow picture of rationality employed in the prisoners' dilemma framework. This is the main reason for which I think Gauthier's attempt fails by his own lights.

CONCLUSION

Gauthier's project, however, is not an insignificant one for my purposes. I think that it, like the iterative project before it, may be understood as part of a redundancy charge. Rather than solving the prisoners' dilemma, Gauthier's account and the iterative story highlight reasons to think that the prisoners' dilemma is irrelevant to the question of public goods cooperation in the real world. Because of certain empirical evidence (in numerous situations people do act on joint strategies, indeed ignoring the payoffs that would be available to them should they defect), I suggest that we don't find a strategy like CM to be intuitively irrational. That it be deemed so when viewed through the lens of the prisoners' dilemma, then, is merely a reason to write off the game in our discourse.

The issue boils down to one of realism. The prisoners' dilemma may be a useful tool in some domains, but it simply doesn't seem to be the case that it accurately represents many real life interactions. Firstly, as suggested in chapter 3, our interactions with others are generally sustained over time, in such a way that the rationality at play within game theory appears to be rather short sited. Secondly, my intuitive reaction to Gauthier's work is that a conditionally cooperative strategy would be far from irrational; there is little reason to think that motivations in real life interactions are restricted to the narrow maximisation that the prisoners' dilemma employs.

All this, of course, goes some way to undermining the economic justification for the state. If, as I suggested, the iteration argument does seem to point to the possibility of cooperation in an iterated game, the preferences of which are those of a prisoners' dilemma, then the pragmatic argument for the state put forward in chapter 2 looks unconvincing. Public goods scenarios, it seems plausible to say, seem to be interactions that (at least have the potential to) stretch into the future; given that these are important goods for individuals, it seems unlikely that their provision (or, indeed, non-provision) would be settled in a manner anything like the one-shot game.

In addition, real life instances of cooperation in which third party enforcement has not proved necessary cast doubt on the notion that individuals could only produce public goods in the presence of the state. The narrow

motivational assumptions at play within the game theoretical framework, it may be argued, are gross over-simplifications of actual human motivations; ruling out any preference for cooperative behaviour as irrational simply seems to conflict with our everyday social experience, in which we often succeed in coming to mutually beneficial arrangements. The suggestion, then, is that, whatever one takes the lesson of the prisoners' dilemma to be, the game doesn't represent a great number of the actual situations we find ourselves in.

The above observations, then, throw doubt on the claim that individuals could not produce public goods without the kind of coercive apparatus provided by the state. As such, they provide reasons to doubt the plausibility of the economic justification for the state. If cooperative behaviour could feasibly arise in the absence of the kind of coercive monopoly that the state constitutes, then we might have to look elsewhere to find pragmatic reasons for accepting it, and the costs it imposes upon us.

It is of course possible that no such reasons exist; that states, by and large, do not represent a prudential 'good deal' for those over whom they claim authority, and are therefore not cannot be justified. If this were the case, then our relationships with our states would be rather worrying, especially if we are sympathetic to the philosophical anarchist's arguments concerning legitimacy.

It was suggested earlier that the justification of states could provide reason to eschew a more overt political anarchism, even if one denied the existence of general political obligation. Even an illegitimate state, I posited, could be deserving of the acceptance or support of its putative citizens, if they judged it to represent a prudentially beneficial arrangement. At the very least, such justification would seem to provide a reason not to engage in action intended to damage that structure. If no such justification were forthcoming – that is, if neither of the distinct manners of normatively assessing of the state speaks in favour of its existence – then this could warrant an important change in our practical attitudes towards the state.

BIBLIOGRAPHY

- Axelrod, Robert. *The Evolution of Cooperation - Revised Edition* (Basic Books, 2006).

- De Jasay, Anthony. 'Self-contradictory contractarianism,' in Sanders/Narveson ed. *For and Against the State* (Rowman and Littlefield, 1996).

- Flathman, Richard E. *The Practice of Political Authority* (University of Chicago Press, 1980).

- Gauthier, David. *Morals by Agreement* (Oxford University Press, 1986).

- Harriott, Howard H. 'Games, Anarchy, and the Nonnecessity of the State,' in Sanders/Narveson ed. *For and Against the State* (Rowman and Littlefield, 1996).

- Hart, H.L.A. 'Are There Any Natural Rights?' in *The Philosophical Review*, Vol. 64, No. 2 (April 1955).

- Luce, R. Duncan & Raiffa, Howard. *Games and Decisions* (New York: Wiley, 1957).

- Narveson, Jan. 'The Anarchist's Case,' in Sanders/Narveson ed. *For and Against the State* (Rowman and Littlefield, 1996).

- Nozick, Robert. *Anarchy, State, and Utopia* (New York: Basic Books, 1974).

- Rainbolt, George W. 'Gauthier on Cooperating in Prisoners' Dilemmas,' in *Analysis*, Vol. 49, No. 4 (October 1989).

- Raz, Joseph. 'Authority and Consent,' in *Virginia Law Review*, Vol.67: 103 (1981).

- Raz, Joseph. 'The Problem of Authority: Revisiting the Service Conception,' in *Minnesota Law Review*, Vol.90, No.4 (April 2006).

- Sartorius, Rolf. 'Political Authority and Political Obligation,' in *Virginia Law Review* Vol. 67, No. 1, The Symposium in Honor of A. D. Woozley: Law and Obedience (February 1981).

- Schmidtz, David. 'Justifying the State,' in Sanders/Narveson ed. *For and Against the State* (Rowman and Littlefield, 1996).

- Sen, Amartya. *Rationality and Freedom* (Belknap: Harvard, 2002).

- Simmons, A. John. *Moral Principles and Political Obligations* (Princeton University Press, 1979).

– Simmons, A.John. 'Philosophical Anarchism,' in Sanders/Narveson ed. *For and Against the State* (Rowman and Littlefield, 1996).

– Simmons, A. John. 'Justification and Legitimacy,' in *Ethics*, Vol. 109, No. 4 (July 1999).

– Simmons, A. John. 'The Duty to Obey and Our Natural Moral Duties,' in Wellman & Simmons, *Is There a Duty to Obey the Law?* (Cambridge University Press, 2005).

– Smith, Holly. 'Deriving Morality from Rationality,' in Vallentyne ed. *Contractarianism and Rational Choice* (Cambridge University Press, 1991).

– Smith, M.B.E. 'Is There a Prima Facie Obligation to Obey the Law?' in *The Yale Law Journal* Vol. 82, No. 5 (April 1973).

– Taylor, Michael. *Community, Anarchy and Liberty* (Cambridge University Press, 1982).

– Taylor, Michael. *The Possibility of Cooperation* (Cambridge University Press, 1987).

– Weber, Max. 'Politics as a Vocation,' in *The Vocation Lectures*, trans. Rodney Livingstone (Hackett, 2004).

– Wolff, Robert Paul. *In Defence of Anarchism* (University of California Press, 1998).