Can Interdependent Reasoning Help Provide a Solution to the Problem of Political Obligation?

George Turturea

Submitted to the Central European University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy
Supervisor: Professor János Kis

Department of Philosophy
Central European University, Budapest
2011

I hereby declare that the dissertation contains no materials accepted for any other degrees in any other institutions of higher education. Also, I declare that the dissertation contains no materials previously written and/or published by another person, except where appropriate acknowledgment is made in the form of bibliographical references.

George Turturea
October 2011

Abstract

My main purpose in this dissertation is to provide an answer to the question whether interdependent reasons for action can help provide a solution to the problem of political obligation. To attain it, I propose an investigation which unfolds in two parts.

In the first part, I examine the relationships that hold between personal autonomy, political authority and rationality, focusing on a reading of these relationships according to which, even though autonomous individuals may not a have reason to comply with the directives of political authority as long as they consider the matter from the standpoint of individual rationality, they may nevertheless have a reason to comply with such directives if they consider the matter from the standpoint of collective rationality. The ensuing attempt to explore the strengths and limitations of a model that tries to provide the normative foundations of political obligation by employing the principle of collective rationality leads me to the conclusion that, while the principle of collective rationality plainly dictates a greater level of compliance with the law than the principle of individual rationality taken in conjunction with an individual's moral values would dictate, the very same considerations which are invoked in order to establish the need to appeal to a principle that would give one a reason to obey the law over and above what substantive moral principles working within the framework of individual rationality give one reason to do might render implausible any attempt to account for political obligation by reference to such substantive moral principles. In order to deal with this difficulty, I suggest that, by introducing a particular conventionalist account, one may take a significant step in the direction of rehabilitating the idea that the reasons for obeying the law can be accounted for by appeal to substantive moral considerations.

In the second part of the dissertation, therefore, I scrutinize the various arguments that the version of a conventionalist account that I focus on can put forward in order to back up the claims that political obligation, if it exists, can only be accounted for in terms of interdependent reasons for action. I first look at the idea that a promising strategy in dealing with the problem of political obligation is to show that the treatment of ordinary coordination problems can be extended to other game theoretic situations, such as the prisoner's dilemma. In this sense, I find that the conventionalist account is able to answer the objection that any conventionalist account of authority and political obligation is misguided because the conventionalist analysis cannot properly be applied to situations of significant conflict of interests. I also argue for the point that, in order to defend the conventionalist view according to which political obligation is to be justified by reference to a pattern of mutual expectations, one should make crucial use of the idea of cooperative dispositions. To this end, I show that conventionalists are right in claiming that assumptions about cooperative dispositions are not *ad hoc*, and, hence, that conventionalism can make sense of the idea that patterns of mutual expectations are not immune to moral criticism.

Acknowledgments

I wish to thank first of all Professor János Kis, my supervisor at CEU, who has been a source of philosophical and personal inspiration for me for so many years. That this dissertation exists at all is primarily due to his patient academic guidance and constant support. I would also like to thank Dr. Daniel McDermott, my supervisor at Oxford. His feedback helped me clarify many ideas which now figure as parts of the first four chapters of my dissertation, and his encouraging and trustful academic supervision greatly contributed to making my stay at Oxford an enjoyable experience.

I also wish to thank my colleagues and friends Andrei Apostol, Cristian and Miha Constantinescu, Anca Gheaus, Serban Popescu-Szekely, Csaba Szilagyi, Adrian Tudorica and Camil Ungureanu, for their kind suggestions, pointed criticisms and amazing camaraderie. For reasons too numerous and varied to detail, I lovingly thank Laura Vasile, my better half.

I am very grateful to Kriszta Biber, our wonderful department coordinator, for helping me in a myriad ways throughout my stay at CEU.

Finally, I would like to gratefully acknowledge the support that I received throughout my doctoral studies from the following institutions: the Central European University in Budapest (my fantastic home university), the Lincoln College and the Department of Philosophy at the University of Oxford (where I spent one academic year as doctoral researcher), the Open Society Archives in Budapest (which awarded me a joint CEU/OSA junior researcher grant), and the United Kingdom's Foreign and Commonwealth Office (for providing the Chevening Scholarship that enabled me to pursue research at the University of Oxford).

Table of Contents

Chapter 1
Introduction

1.1 Accounting for political obligation in terms of interdependent reasons for action

According to T. H. Green's rightfully celebrated view, the primary error of the way in which classical political theories deal with the problem of political obligation lies in the manner in which they phrase the very question they are supposed to answer. These theories, Green writes:

> […] make no inquiry into the development of society and of man through society. […] They leave out of sight the process by which men have been clothed with rights and duties, and with senses of right and duty, which are neither natural nor derived from a sovereign power. They look only to the supreme coercive power on the one side and to individuals, to whom natural rights are ascribed, on the other, and ask what is the nature and origin of the right of that supreme coercive power as against these natural rights of individuals. The question so put can only be answered by some device for representing the individuals governed as consenting parties to the exercise of government over them. […] But in truth it is only as members of a society, as recognizing common interests and objects, that individuals come to have these attributes and rights.[1]

As analyzed by Green, then, such political theories as those of Hobbes, Locke and Rousseau, primarily concerned as they are with the relationship between, on the one hand, the individual endowed with natural rights and liberties and, on the other hand, the political authority, are fraught with difficulties primarily due to the fact that they misrepresent this relationship as one between two isolated poles, with nothing but a "fictitious explanation" left to account for its nature.

[1] T. H. Green, *Lectures on the Principles of Political Obligation*, pp. 121–122. Note that for the purposes of the present introduction by political obligation I will simply mean the moral obligation that the individual has to obey the law *qua* law.

In one form or another, views based on Green's intuition have been deployed by many political philosophers since his time.[2] One such view is an elaborated contemporary theory which reads Green as focusing on the implausibility of viewing morality as an eminently individual affair. Gerald Gaus, the proponent of this reading, understands Green as pointing to the difficulties that derive from assuming that "there is nothing between the private individual conscience and the political-moral authority of the state".[3] The major problem with casting the relationship between the individual and the political authority in such terms is the over-individualized conception of morality which necessarily accompanies it. This is an undesirable fellowship, according to Gaus, for at least two reasons. For one, it undermines the moral side of the relationship, by rendering it unable to accomplish a task that any proper moral theory should be able to fulfill, namely that of accounting for the fact that there is a strongly social streak to morality. Furthermore, it undermines the political side of the relationship, by making the resulting conception of authority vulnerable to the claim that subjection to a political authority conceived in its terms will necessarily be in conflict with basic tenets of individual morality. Gaus finds implausible both the idea that the dictates of morality can appropriately be viewed as simply the dictates of an individual's conscience, and the related idea that political authority should be resisted because subjecting oneself to it would go against the dictates of private conscience.

[2] On a very general level, one could argue that political theories which emphasize the pivotal role that the political community should play seek to substantiate an intuition similar to Green's, while not many of them share (nor, indeed, need to share) Green's further theoretical commitments, e.g. the organicist character of his views. For an excellent recent discussion of Green's political theory, see David Brink's *Perfectionism and the Common Good: Themes in the Philosophy of T.H. Green*.

[3] See Gerald Gaus, *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, p. 450.

These brief remarks are intended to illustrate one of the fundamental intuitions upon which much of this dissertation is built. The thought is that a strict bilateral way of conceiving the relationship between the individual and the political authority is likely to lead to an unjustified rebuttal of both valid ways of grounding the legitimacy of political authority and plausible ways of justifying political obligations. One of my aims will be, then, to look at what I take to be the most plausible arguments for the claim that a proper account of political obligation needs to go beyond viewing the individual as essentially locked into an antagonistic relationship with the political authority. To this effect, I will consider theories that place a special emphasis on the interdependent character of the reasons for actions upon which political obligation can arguably be grounded.

I begin somewhat indirectly, by looking at a view which I take to be fully committed to a contemporary version of the opposition that T. H. Green had in mind. It is a view which, precisely due to the over-individualized notion of the demands of rational and moral agency which it employs, ends up by putting forward an implausible set of theoretical commitments. According to it, a proper normative understanding of personal autonomy is one which is necessarily conducive to the idea that subjecting oneself to political authority amounts to a betrayal of one's autonomy. On this view, there is an inherent tension between the proper exercise of autonomy and the acceptance of the directives of political authority as morally binding. It would then seem that, on pain of sheer heteronomy, the only way out of this tension is to adopt philosophical anarchism.

The advocate of this philosophical outlook claims that the idea of personal autonomy must be understood in the sense that the individual has a moral duty to be

autonomous and, furthermore, that individuals must form, pass and act on their own judgments on moral matters in order to discharge this duty. However, by accepting the directives of political authority as reasons for action, individuals would surrender their private judgment and, consequently, forfeit their autonomy, because to accept a state's claim to authority means to recognize the duty to obey its commands simply by virtue of the fact that they are issued by that state. This view of personal autonomy is rightfully criticized on the grounds that the notion of a moral duty to be autonomous that it employs is implausibly strong. As it has been observed, however, even if one gives up this moralized view of autonomy (e.g. by preferring to understand it as a capacity or an ideal of self-governance to be realized by listening to the various types of dictates of the individual conscience), one is left with the difficulty of explaining how is it possible for the individual to always accept as *rationally* binding the dictates of an authority which will at least sometimes direct him to act contrary to the balance of reasons as he sees them.[4] In other words, one still has to provide an adequate answer to the charge that it is simply irrational to surrender one's judgment on the balance of reasons, in order to act on the directives of authority.

Within this framework, I proceed by looking at what is often regarded as the most plausible understanding of authority that one can employ in order to successfully refute the arguments advanced by philosophical anarchists – the service conception put forward by Joseph Raz. I do this with an eye at showing how some of the most problematic consequences that result from conceiving political authority along these lines impact both on the relationship between authority and rationality, and on the way in which this

[4] See Leslie Green, *The Authority of the State*, pp. 25–26.

relationship is relevant for an individual's reasoned decision whether to cooperate with the political authority.

A substantial part of the dissertation is devoted to analyzing a recent attempt to justify obedience to the law and political authority by reference to the idea of collective rationality. A promising response to the challenge posed by philosophical anarchism is to argue that, even though individuals do not have a reason to comply with the law as long as they consider the matter from the standpoint of individual rationality, they might nevertheless have a reason to comply with it if they consider the matter from the standpoint of collective rationality. The proponent of this view, Christopher McMahon, points out that once the possibility of acting as a free rider is taken into account, each may judge that he would be better off, in light of his own values, by refusing to subject himself to the directives of authority. However, given that each may find himself in a multi-person prisoner's dilemma situation, the reverse might the case, i.e. each may judge that the values he regards as applicable would be better served by following the directives of authority. The principle of collective rationality gives one a reason to comply with the law by directing one to compare, from the standpoint of one's values, the cooperative outcome with the best one could achieve in the noncooperative outcome, i.e. the outcome in which everyone chooses to defect.

According to McMahon, the rationality of contributing to a cooperative scheme (e.g. the rationality of contributing to the maintenance of a state by obeying its laws) is accounted for by the fact that the cooperative outcome is preferable, from the standpoint of one's values, to the best one would be able to achieve if the cooperative scheme did

not exist. On his view, the need to account for the reason to contribute to mutually beneficial cooperative schemes by invoking the principle of collective rationality stems from the fact that traditional attempts to account for this reason by invoking substantive moral principles (e.g. the principle of fairness, the principle of consent or the natural duty of justice) fall prey to a common objection. These principles, the objection goes, cannot reliably block the threat of morally motivated defection. More specifically, one can point out that, when individuals have moral reasons for defecting (disobeying the law), the moral reason that supports contribution must compete with the moral reason that supports defection, and there is no guarantee that the former will override the latter. Admittedly, an individual might be able to achieve a great deal in terms of his values by putting to a different use the resources he is called upon to contribute, and so, he might judge that defecting from the scheme is morally superior to contributing. At this point, many would insist that a fairness-based approach to cooperation has the resources to solve the problem of morally motivated defection. The idea is that individuals who care about fairness would always judge that the reason of fairness for contributing prevails over moral reasons for defecting. This line of argument, however, is open to a serious objection. As McMahon points out, to accept that one's moral concerns can appropriately be sacrificed in order to satisfy the requirement of fairness means to be prepared to demote one's moral concerns to the status of interests. The principle of collective rationality can arguably avoid both the above-mentioned problems. It reliably blocks the threat of morally motivated defection by providing one with a new reason to contribute, a reason that one would not have if one considered what substantive moral considerations in conjunction with the principle of individual rationality require. The principle of collective

rationality gives one sufficient reason to contribute to a scheme if one judges the cooperative outcome to be preferable to the outcome in which everyone defects. Equally importantly, it prompts one to compare the two outcomes from the standpoint of one's undemoted values.

As it will turn out, the adoption of the principle of collective rationality as a characterization of the requirement to contribute to cooperative schemes can be viewed as an attempt to substantiate the intuition that, if it is to be justified, this requirement must be couched in terms of interdependent reasons for action. The principle of collective rationality justifies contribution to a cooperative scheme only insofar as the assurance problem is solved (i.e. it gives each sufficient reason to contribute only insofar as there are reasons to believe that enough others will contribute so that the result will be an outcome that each can regard as preferable to the noncooperative outcome). Furthermore, unlike the principle of individual rationality, which directs the agent considering whether to contribute to a cooperative scheme to compare the outcomes produced by contributing and defecting, taken as individual events, the principle of collective rationality directs the agent to compare the cooperative outcome in which one's contribution is added to the contributions that are actually made by others with the noncooperative outcome in which everyone would defect.

My aim in this part of the dissertation is to explore the strengths and limitations of a model that attempts to provide the normative foundations of political obligation by employing the principle of collective rationality. The main worry is that, while the principle of collective rationality might ultimately fail to ground a suitably general obligation to obey the law, the considerations that are invoked to establish in the first

place the need to appeal to a principle that would give one a reason to obey the law over and above what moral principles that work within the framework of individual rationality give one reason to do, might render implausible any attempt to account for political obligation in terms of such substantive moral principles. In other words, given that morally motivated defection poses a serious problem for any such attempt, and that invoking fairness in adjudicating between conflicting moral points view seems to lead to a problem that is even deeper than the one it solves (i.e. the demotion of moral concerns to interests), those who want to justify political obligation by reference to substantive moral principles would have provide an argument to the effect that McMahon's treatment of the above-mentioned issues is wanting. One of my aims will be to offer a brief outline of the strategy that can be adopted by those who take seriously McMahon's objections to characterizing the requirement to contribute to cooperative schemes in terms of substantive moral principles but who nevertheless consider that employing such principles is preferable to employing a formal principle such as the principle of collective rationality. As it will turn out, the resources provided by a conventionalist account may help take a significant step in the direction of rehabilitating the idea that the reason to contribute to cooperative schemes is best accounted for by appeal to substantive moral principles.

Next, I look at another attempt to account for political obligation in terms of interdependent reasons for action. More specifically, while the broader aim of the ensuing discussion is that of assessing the claim that, as a normative framework, conventionalism can provide a valuable insight into the problem of justifying both political authority and

political obligation, I begin by analyzing how contemporary conventionalist accounts inspired by the work of David Lewis can constitute plausible venues in the search for interdependent reasons for action.

Famously defended by David Hume, conventionalism has an enviable pedigree among normative theories. While one could even claim that, under the guise of positivism, it is one type or another of normative conventionalism that presently enjoys a position of dominance within the realm of legal theory, within the realm of political philosophy conventionalism does not seem to be kept in a similarly high regard, despite the fact that even staunch anti-conventionalists seem to agree that this position presents certain theoretical advantages.[5] One of the most plausible explanations of this dissimilarity is that, since political philosophers who attempt to account for the obligation to obey the law are looking for moral reasons for obedience, they quite understandably dismiss conventionalism as unable to provide such reasons due to its being a normative framework which is purely instrumentalist in character.

A promising strategy to highlight the specific contribution of conventionalism to solving the problem of political obligation is to make the case that, despite its instrumentalist character, conventionalism is a necessary element of the normative background upon which any plausible derivation of the reasons to obey the law from substantive moral principles must rest. In this sense, one can argue that, due to its emphasis on the study of coordination problems and of the conditions for the emergence of patterns of mutual expectations, conventionalism is uniquely well-placed to play the

[5] Leslie Green, for instance, admits that conventions constitute an interesting venue for the study of political authority both historically (because they represent an argument often invoked) and theoretically (because, since they lack the problematic aspect of enforcement, they make the search for an indirect justification of authority a lot easier). See Green, *The Authority of the State*, chapter 4.

role of a constitutive element of such a normative background. As I already suggested, an argument to this effect could be bolstered by offering a detailed discussion of David Lewis's view of conventions, given that, while working within the confines of coordination games, Lewis defines a convention as a regularity in behaviour which holds – and it is commonly known to hold - within a given group confronted with a recurrent situation, thus paving the way for a swift move to the study of the patterns of mutual expectations that emerge by observing conventions. In studying such patterns of expectations, however, one should not only analyze the way they function (i.e. the internal coherence of the process that begins with their emergence and ends with their acquiring normative powers), but should also deal with the inherent limitations of a theoretical model which centres exclusively on them.

My own discussion will focus on whether such limitations can be superseded by going beyond the Lewisian understanding of conventionalism towards a more comprehensive view which, starting from certain assumptions about the cooperative virtues that can plausibly be ascribed to people, highlights both the interdependent character of reasons for action and the necessary interplay between the relevant patterns of mutual expectations and particular substantive moral principles. If what one attempts to do is to outline a plausible theory of political obligation, the need to move beyond Lewisian conventionalism becomes straightforward, since Lewis's theory deals only with coordination problems. Pace those authors who consider that coordination games constitute the preeminent type of game according to which the normative analysis of political societies should be modelled[6], any effort to model the interactions within large

[6] Jeremy Waldron claims that "for the purposes of normative analysis, partial-conflict coordination games capture the essence of politics" (Jeremy Waldron, *Authority for Officials*, p. 51, n. 20). Joseph Raz also

scale political societies would gain in plausibility by considering strategic games other than coordination games. At this point, it is worth noting that McMahon shares with conventionalists not only the idea that political obligation, if it exists, can only be accounted for in terms of interdependent reasons for action, but also the idea that a promising strategy in dealing with the problem of political obligation would be to find a way of extending the treatment of ordinary coordination problems to multi-person prisoner's dilemmas. On the one hand, according to McMahon, what enables cooperatively disposed individuals to view prisoner's dilemma situations as having the structure of coordination problems is the principle of collective rationality (under this principle individuals assign the same payoff to unilateral defection that they assign to the noncooperative outcome). On the other hand, according to certain proponents of conventionalism, the rules that arise in patterns of interaction that take the form of prisoner's dilemma have themselves a similar structure to conventions that govern coordination games (i.e. they are patterns of mutual expectations concerning individual decisions).[7] This being so, one of my aims will be to see whether the latter account is preferable to the former, i.e. whether it can contribute to solving the problem of political obligation in ways in which the former account cannot.

stresses the importance of "the need to co-ordinate the action of several people", claiming that "All political authority rests on this foundation (though not only on it)" (Joseph Raz, *Practical Reason and Norms*, p. 64). For further comments on these points, see Leslie Green, *The Authority of the State*, p. 109, and *Strategy and Ultimate Legal Rules*, pp. 69–74.

[7] See Govert den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 216.

1.2 Overview

In Chapter 2, I examine the interaction between autonomy, authority and rationality, advancing a sum of considerations on the relevance of undertaking the project that I develop throughout the dissertation. The personal autonomy of the moral agent seems to be in a potentially conflicting relation with various types of external authority, of which political authority is a distinct species. The same seems to be true about the character of the relation between personal autonomy and any project couched in terms of collective rationality, at least if we believe that being autonomous necessarily implies keeping to the strong preeminence of the personal point of view in reaching practical decisions, while the idea of 'collective rationality' seems to intuitively lead us in a somehow opposite direction, whatever that direction might involve. I begin the chapter with a sketchy look at the relevant way in which, despite their common origin, personal autonomy differs from, and is constrained by what might be provisionally called 'autonomy as manifested in politics'. I then go on to consider certain aspects of philosophical anarchism, a theoretical stance advanced (implausibly, as I will argue) as a distinctive way out of the tension between autonomy and authority. Following up on these developments, in the last section I argue for the claim that a proper view of autonomy should stick to the requirements of rationality, and, via a short reprise of the issue of philosophical anarchism, I outline the main direction of Chapter 3, which focuses on (political) authority.

In Chapter 3, I place under critical scrutiny Joseph Raz's model of authority, which is widely regarded as the most plausible model for the justification of political authority available, as well as argue for the need for adopting a model of collective

rationality, a need resulted both from the necessity of finding a mediating factor in the announced conflict between authority and autonomy, and from the inability of Raz's model to justify political obligation. A scrutiny of the proposed framework is put forward in the last section of chapter 3, in an attempt to outline one of the important directions of the second part of the thesis.

In Chapter 4, I analyze the suggestion developed by Christopher McMahon according to which a formal principle such as the principle of collective rationality is preferable to a substantive moral principle as a characterization of the requirement to contribute to mutually beneficial cooperative schemes. More specifically, I consider McMahon's argument to the effect that, unlike the moral principles traditionally invoked to account for the aforementioned requirement, the principle of collective rationality can successfully block the threat of morally motivated defection. Furthermore, I take a close look at his claim that any attempt to characterize the requirement to contribute to cooperative schemes by appeal to the principle of fairness is objectionable because it implies that individuals should be prepared to demote their moral concerns to interests.

I then move on to spell out the implications of McMahon's account of collective rationality for the problem of political obligation. While the principle of collective rationality apparently justifies a greater a level of compliance with the law than the moral principles which work within the framework of individual rationality can justify, it is doubtful that this principle can ground a suitably general obligation to comply the law. I argue that, despite its initial appeal, McMahon's argument that substantive moral principles cannot play the central role typically envisaged for them within a theory of political obligation fails. As it will turn out, his defense of the idea that the principle of

fairness cannot offer an adequate characterization of the requirement to contribute to cooperative schemes because fairness cannot appropriately govern the realization of individuals' moral concerns trades on an equivocation between different types of cases in which the principle of fairness would be invoked to adjudicate between conflicting moral points of view. Moreover, I will argue that whether one accepts McMahon's view that the principle of collective rationality is needed to justify contribution to cooperative schemes will depend on whether one is skeptical about the possibility of choosing a scheme that each can regard as justified. Yet, if an approach to cooperation which attempts to prove that the members of a group can choose a scheme that all can deem justified is vindicated (for instance, because evaluative standards themselves can transform in light of the moral benefits produced by cooperation), the principle of collective rationality can be dispensed with.

In Chapter 5, I look at the merits of contemporary conventionalism and attempt to assess the claim that, as a theoretical framework, conventionalism has valuable resources to provide an innovative insight into both the problem of justifying legal and political authority and the problem of political obligation. I begin by offering a picture of the way in which conventionalism enters contemporary debates. In the following sections, I proceed by looking at the arguments put forward by David Lewis in support of his own version of conventionalism, and at a certain understanding of conventionalism which draws upon his account. One of the upshots of the arguments in these sections is that a non-arbitrary understanding of convention is indeed plausible. I then proceed by looking at the main claims that should characterize a plausible conventionalist account of authority. The considerations presented in this chapter lead me towards an analysis of a

further claim, namely that a proper characterization of political obligations will necessarily include an appeal to the presupposition that people are cooperatively disposed. Finally, I argue for the plausibility of an understanding of political obligation developed within a conventionalist framework qualified in such a way as to make room for the normative use of substantive moral principles such as the principle of fairness.

As already emphasized, both McMahon's attempt to ground political obligation on the principle of collective rationality and the conventionalist attempt to account for political obligation are guided by the conviction that, if there are reasons to obey the law, such reasons can only be interdependent ones. Yet, even though they find implausible the view of agency according to which the individual is an isolated rational actor, both McMahon and conventionalists argue for the importance of holding on to a notion of individual, as opposed to collective, agency (they argue, for instance against interpreting mutually adjusted action by reference to a strong concept of collective intention). In the Appendix, I turn to one of the better known attempts to defend a robust conception of collective agency and discuss some of the difficulties involved in applying it to the problem of political obligation. More specifically, my aim is to offer a critical analysis of Margaret Gilbert's plural subject theory of political obligation. My analysis attempts to answer three questions. First, I focus on the question whether Gilbert is right in asserting that joint commitments establish obligations and entitlements. The second question that I attempt to answer is whether obligations and entitlements derive from joint commitments irrespective of any other considerations. More concretely, I consider the question whether there are sufficient reasons for accepting that being obligated by a joint commitment is not precluded by coercive circumstances or coercive content. Finally, I attempt to assess

the difficulties involved in applying Gilbert's general account of joint commitments to the political sphere in a way that yields political obligations.

Chapter 2
The interplay between autonomy, authority and rationality

In this chapter, I place under scrutiny the interaction between autonomy, authority and rationality. The personal autonomy of the moral agent seems to be in a potentially conflicting relation with various types of external authority, of which political authority is a distinct species. The same seems to be true about the character of the relation between personal autonomy and any project couched in terms of collective rationality, at least if we believe that being autonomous necessarily implies keeping to the strong preeminence of the personal point of view in reaching practical decisions, while the idea of 'collective rationality' seems to intuitively lead us in a somehow opposite direction, whatever that direction might involve. I begin the chapter with a sketchy look at the relevant way in which, despite their common origin, personal autonomy differs from, and is constrained by what might be provisionally called 'autonomy as manifested in politics'. In the second section, I consider certain aspects of philosophical anarchism, a theoretical stance advanced (implausibly, as I will argue) as a distinctive way out of the tension between autonomy and authority. Following up on the developments from this section, in the last section I argue for the claim that a proper view of autonomy should stick to the requirements of rationality, and, via a short reprise of the issue of philosophical anarchism, I outline the main direction of the second chapter, which focuses on (political) authority.

2.1 Autonomy and rationality

The idea of personal autonomy springs from a parallel with the idea of group autonomy. This parallel is emphasized by Joel Feinberg, who, after indicating that autonomy has the sense of "self-rule", "self-determination" self-government" and "independence", argues that "it is plausible that the original applications and denials of these notions were to state, and that their attribution to individuals is derivative, in which case 'personal autonomy' is a political metaphor".[8]

The point that the notion of personal autonomy originates in a political metaphor is also stressed is also stressed by Sarah Buss. According to her, when a group of people make reference to their right to live autonomously they mean "that they ought to be allowed to govern themselves". In doing so, those in the group "are, in essence, rejecting the political and legal authority of those not in their group. They are insisting that whatever *power* these outsiders may have over them, this power is illegitimate; they, and they alone, have the *authority* to determine and enforce the rules and policies that govern their lives." Thus, the new spin is that the proper parallel to be drawn between a group's claim to autonomy and an individual's claim to be autonomous within a certain sphere of life implies that the latter claim too is necessarily connected with the denial of anyone else's authority to control that individual's activity within the sphere in question. This implies saying, in fact, that "any exercise of power over this activity is illegitimate unless she authorizes it herself".

Buss's account is important insofar as it stresses the existence of one very important exception to this parallelism, a reason that applies only to 'personal autonomy'

[8] Joel Feinberg, *Harm to Self*, pp. 27–28.

and that apparently takes us beyond politics. This reason boils down to the claim that, insofar as one is an agent, one is correct to regard one's own commitments, judgments and decisions about how one should act, as *authoritative*. If one were to challenge the authority that is an essential feature of one's judgments and decisions, then they would cease to be one's own practical conclusions. This is so because, as Buss writes,

> […] every agent has an authority over herself that is grounded, not in her political or social role, nor in any law or custom, but in the simple fact that she alone can initiate her actions. (…) The point is that she has no conceivable option. In order to form an intention to do one thing rather than another, an agent must regard her own judgment about how to act as authoritative.[9]

This move, however, opens up a discussion about personal autonomy and the contours of human (or, more specifically, moral) agency that, as such, is often seen as quite separate from the usual ways of tackling with autonomy within the realm of political philosophy. This aspect is of particular importance: if a sharp dividing line between personal and political autonomy exists, then the idea of autonomy cannot be embodied in a unitary concept. Consequently, two major questions arise. First, are there any features that constitute a minimal core of the idea of autonomy, features which must be preserved in both concepts (i.e. 'personal' and 'political' autonomy)? Second, how should we understand the contention that there is a special meaning of personal autonomy that should/can be properly used in politics, and how must this meaning be distinguished from other important senses that the concept of personal autonomy, possesses?

Contemporary philosophers writing in the liberal tradition seem to be on opposite sides of this debate, although they do not seem to see it as impossible to overcome. An

[9] Sarah Buss, *Personal Autonomy.*

important theorist of autonomy like Harry Frankfurt, for instance, dedicates an entire section of a paper aimed at refuting the philosophical anarchism put forward by Robert Paul Wolff to arguing for the claim that "Wolff misconstrues the extent to which the notion of autonomy is pertinent to the analysis of political relationships."[10] What he seems to imply is that, while autonomy properly understood is a unitary concept, some of its applications are improper. By contrast, Joseph Raz warns that "personal autonomy, which is a particular ideal of individual well-being, should not be confused with the only very indirectly related notion of moral autonomy". This is so, Raz claims, because personal autonomy "is essentially about the freedom of persons to choose their own life" which, if valid, can only figure as "one element in a moral doctrine", while moral autonomy, originated in Kant's view that morality consists of self-enacted principles, "is a doctrine about the nature of morality".

Finally, John Rawls also implies, in his later work, that ethical autonomy is distinct from political autonomy. The intricate character of his discussion of autonomy is partly due to the fact that, in *Political Liberalism*, he further distinguishes between rational autonomy and full autonomy. The divide between the latter two notions marks, in a certain sense, a continuum. As he writes,

> Citizens think of themselves as free in three respects: first, as having the moral power to form, to revise, and rationally pursue a conception of the good; second, as being self-authenticated sources of valid claims; and third, as capable of taking responsibility for their ends. Being free in these respects enables citizens to be both rationally and fully autonomous.[11]

[10] H. Frankfurt, *The Anarchism of Robert Paul Wolff*, pp. 410–411.
[11] Rawls, *Political Liberalism*, p. 72.

The first two of the respects in which citizens think of themselves as free inform rational autonomy. The third is necessarily involved in the move from rational to full autonomy, move which is equivalent with the one from the original position to a well-ordered society structured according to the tenets of political liberalism. Out of them, the second respect (i.e. the one that consists in citizens' viewing themselves as being self-authenticated sources of valid claims) seems to come close to Buss's claim that in order to form an intention to do one thing rather than another, an agent must regard her own judgment about how to act as authoritative.

This is a complicated claim to make, however, if we recall that, for Rawls, one of the special features of the political relationship in a constitutional regime is that "it is a relationship of persons within the basic structure of society, a structure of basic institutions we enter by birth and exit only by death (or so we may appropriately assume)." On this point, Rawls writes further:

> To us it seems that we have simply materialized, as it were, from nowhere at this position in this social world with all its advantages and disadvantages, according to our good or bad fortune. I say from nowhere because we have no prior public or nonpublic identity: we have not come from somewhere else into this social world. Political society is closed: we come to be within it and we do not, and indeed cannot, enter or leave it voluntarily.[12]

The complications that I have in mind can be summarized in the question 'how can there be any place for the equal autonomy of all persons to evaluate and revise the structure of, say, their plans and commitments, both personal and political, when they in fact stand under the inexorable limits placed by the unavoidable character of their being a part of a political society'? Attempting to answer this question within Rawls's framework

[12] Rawls, *Political Liberalism*, pp. 135–136.

would presuppose structuring this entire thesis in light of his overall view of political liberalism. Since I do not intend to take this path, however rewarding it might prove to be, let me move on by noting that a hint at the type of answer that might be need to answer this question is suggested by noticing that similar kinds of concerns and complications arise with respect to the idea that there are limits on one's competence to act on one's own reasons. They spring, partly, from our regard for the moral agents' rationality and autonomy. First, one has to consider the question 'how can it be rational to recognize externally imposed limits on one's competence'? Secondly, one must inquire into how can there be any place for the equal autonomy of all persons to evaluate and revise the structure of, say, their moral outlook when they in fact stand under the limits placed by an external authority? The need to provide a plausible answer to the second question becomes especially salient if we recall that these very limits are limits on the occasions on which fundamental revisions can be made. Hence, to stand under an external authority seems to imply that the competence to make such fundamental revisions belongs to someone else (e.g. to an elite, or to the majority through the political conventions they establish). I will try to address both these concerns, in their own turn.

More often than not, personal autonomy is associated with the idea of the rational agent, in the sense that an adequate conception of personal autonomy must require agents to make choices based on critical evaluation of the options before them, and to examine the grounds their beliefs and commitments. However, this idea is resisted by those who claim that if rationality is built into the notion of autonomy, it will obscure the idea of self-government. On this last view, one's government by one's actual self must be

possible regardless of whether one cares about the substantive ideal of rational reflection or not.

Not confusing self-government with the ideal of an examined life, it seems plausible that those who value self-government must also care about acting as they have reasons to act and, moreover, that part of our valuing self-government is due to our valuing the stance of acting for *good* reasons. Why is this so? For one thing, one's feeling a sheer indifference to the reasons for which one acts is an implausible candidate as a lasting psychological attitude. Further, we value agency only partly because we value the activity of rational deliberation and decision. Arguably, an important reason why we bother to make our own choices is that we want to get things right. We care about how our lives go, what sorts of persons we become, and what happens to the various things and persons we care about. If we gave up practical reasoning and choice, the risk would be that perhaps our future behavior would be meaningless, or even harmful. This is why people do not readily give up parts of their agency unless this adds coherence to their lives (e.g. as is the case when we defer our opinion on certain factual matters to experts). Hence, an important reason why we resist having our agency undermined is that we care about the reasons for which we act. Is this too strong a claim? I don't think so. In fact, to be perfectly consistent I should add that our interest in reasons also explains why some causal chains of events/states of affairs do not readily and easily undermine our autonomy: it simply is sometimes rational for us to hold on to values and beliefs which we acquired irrationally.[13]

[13] Note however that, in this case, our interest in reasons is not doing the entire job, but is complemented by the need to preserve one's integrity.

Arguing for this view also might also enable us to understand what is wrong with accounts of autonomy that insist on the agent's substantive independence. On certain plausible assumptions about the social nature of human beings, it seems that few individuals would really add reason to (and increase the overall rationality of) their lives by trying to completely cut themselves off from the social relations and dependencies in which they are embedded.

Those who see autonomy and rationality as normatively severed seem to fail to understand why we care about autonomy in the first place. Their claim is that all one needs for autonomy is one's self and the capacity to express it in action. One lacks autonomy when one acts compulsively, when one is coerced or manipulated, but not when one uncritically takes for granted the values, beliefs and judgments of others instead of forming one's own. An important motivation of this view seems to be the need to avoid bias against certain conceptions of the good. However, this need can be accounted for without severing the link between autonomy and rationality if, for instance, we view autonomy as an ideal which underlies the principle of respect for persons (including, that is, for those who fall short of 'ideal' autonomy). A society should definitely not foster institutions or practices that disregard the preferences or conceptions of the good of some people simply on the grounds that they prone to unreflectively conform to, say, whatever social institutions or practices there might exist in that society. Rather, the question is whether, assuming that every individual member of a political society does envision a certain societal state in which a balance between all the moral considerations available in the considered society obtains, and that their visions of such

societal states will comprise certain sets of morally important social values, it can still be shown that individuals have reasons to cooperate in order to promote such values?

Let me return to the second of the concerns announced above, namely that of inquiring into how can there be any place for the equal autonomy of all persons to evaluate and revise the structure of, say, their moral outlook when they in fact stand under the limits placed by an external authority? To begin to answer this question (and it just this that I attempt to do now, i.e. to sketch out the beginning of an answer), it is important to notice that the promotion of social values will require a collective effort. Hence, there is hope that their promotion will provide both the representatives of the political authority and the rest of its subjects with powerful reasons for action. To better understand why this is so, let me advance three related claims. First, we can plausibly claim that some social values are the kind of goods in the constitution of which a duty is incorporated. In *Liberating duties*, Raz argues that, on the one hand, "duties may be internally related to their justifying goods", and, on the other hand, that there are "some activities and relationships which cannot be specified except by reference to duties", some of these last ones being "intrinsically good".[14] In this sense (i.e. as incorporating a sort of 'duty of promotion'), it can be claimed that the promotion of highly regarded values offers all those who participate in political society reasons for cooperation. Secondly, we can appeal to the common idea of a convergent practice. Since such a practice is usually at the core of a given social rule, if social rules are given normative strength, then some convergent practice must usually figure in the explanation. Note that a converging practice can easily be reason-giving. On the assumption that I am self-interested, the fact that everyone drives on the right side of the road gives me a reason for

[14] Raz, *Liberating Duties*, p. 41.

doing the same thing. My interests can often require me to coordinate my behavior with others, and when this is so, I acquire a reason to do something that others do simply because they do it. If I'm not exclusively self-interested – say I'm motivated to do the right thing but am uncertain about what morality requires of me –, then, if I believe that others are similarly motivated, I have a reason to do what they are doing. This is not primarily so because I have to coordinate with others, but is rather due to the fact that, given certain assumptions about authority, I am more likely to be doing what I ought to do, that is, the right thing by following their lead. Finally, it is also plausible that certain deontological principles present in a given society are observed by all the members of that society, even if directly only in the case of those who uphold them and merely indirectly in the case of, say, public officials, who would strive not to obstruct them. It follows from these considerations that precisely those elements which are constitutive of individuals' conceptions of the public good provide individuals with moral and prudential reasons for cooperation. Also, since in this way the political authority's directives will be based on a conception of the public good whose elements provide reasons for action for its subjects, it can be said that the authority grounds its directives in reasons that already independently apply to its subjects.

Now, instead of inappropriately delving further into these issues at this point in the thesis, I think that it is more suitable to look at a theoretical position according to which any attempt to establish a relation of subjection between individuals and political authorities seems doomed due, importantly, to the intolerable constraints that such subjection would necessarily impose on the autonomy of the individuals. It is, of course, the position embraced by philosophical anarchists.

2.2 Authority and the challenge of philosophical anarchism

In a plausible development of his arguments, it seems unavoidable that the proponent of philosophical anarchism will take the following path: he will concede to a certain degree that individuals cooperate in order to create and maintain political societies, will proceed by construing a necessarily complex, yet contingently frightening image of what it is implied by the idea of political authority (as well as by its practical stance), will proceed further by evoking an image of the individual as torn between his original freedom and political necessity and will end by restating the claim that what is minimally implied in the relationship between individuals and their governments is an evil feature of subordination of their will from the part of the individuals. This 'objectionable minimum' will then be unpacked in a variety of ways, depending on the other main traits of each theorist's position. Most of the times, the advocate of philosophical anarchism will also admit that individuals faced with a government's authoritative directives are supposed to take these very directives, rather than the fact that they are backed by threats, as reasons for action. He will argue, however, that this cannot be the whole story: since claiming authority necessarily leads to advancing a requirement to substitute for an individual's own judgment the judgment of another, the demand - implied by the authoritative directives of a government - that individuals act on the reasons offered by the political authority rather than on their own reasons must be rejected.

For Wolff, the staunchest defender of this view, the defense of anarchism rests upon a preferred account of individual autonomy and its relation to the authority of the

state. His arguments, reminding of those of William Godwin, are exemplary for making the case against relations of authority from the standpoint of an individual's claims of conscience.[15] In Wolff's view, to be sure, even a freely taken decision to substitute for one's own conscience the will of another must be rejected, since it inevitably leads to essentially the same loss of dignity and autonomy as the one that characterizes unqualified obedience. According to Wolff, individuals are endowed with *free will* and *reason*, and hence have the capacity to make responsible choices. Moreover, he tells us, individuals are responsible for their choices and have the primary moral obligation to take full responsibility for their actions. Hence, they have the duty to be morally autonomous, and this must result in 'personalized' decision-making. In order to discharge it, individuals must form, pass and act on their own judgments on moral matters. It turns out that, by accepting authoritative commands as reasons for action, individuals would surrender their private judgment, and consequently, forfeit their autonomy. This is because to accept the state's claim to authority means to recognize the duty to obey its commands simply by virtue of the fact that they are issued by the state. Instead, most that a government should do is to offer individuals advice, but individuals' decision regarding any constraints that can be placed on their actions have to be made considering only those reasons for actions that they already accept. Although conscious of his being bound by moral constraints, Wolff's individual must insist "that he alone is the judge of those constraints. He may listen to the advice of others, but he makes it his own by determining

[15] Godwin's assertion that "there is but one power to which I can yield a heart-felt obedience, the decision of my own understanding, the dictate of my own conscience", is seen by many as providing the motivation for Wolff's conception as well. See Godwin, *Enquiry Concerning Political Justice*, p 229. For a parallel with Wolff, see Leslie Green, *The Authority of the State*, p. 24.

for himself whether it is good advice".[16] This is so due to the need of preserving individual autonomy. It is one of Wolff's strongest claims that the individual is presented with the following dilemma:

> If the individual retains his autonomy by reserving to himself in each instance the final decision whether to cooperate, he thereby denies the authority of the state; if, on the other hand, he submits to the state and accepts its claim to authority then … he loses his autonomy.[17]

Since it is trivially true that mere advice is not a good approximation for what the claims advanced by political authoritative amount to, it follows that the conflict between the autonomy of the individual and a government's authority is unavoidable and insurmountable. Hence, the only solution is to regard political authority as lacking any normative justification, and to maintain that "the concept of a *de jure* legitimate state appears to be vacuous". The upshot of Wolff's argument is that 'anarchism is the only political doctrine consistent with the virtue of autonomy'.[18]

Strong considerations plead against Wolff's claim according to which accepting political authority amounts always to surrendering one's autonomy. According to Wolff, the main reason why authority is incompatible with autonomy is the content-independence of authoritative demands. He rightly stresses the idea that authoritative demands should be distinguished from persuasive arguments.[19] Authority issues binding directives that are to be obeyed irrespective of the subject's judgment on the merits of the directives; in other words, the *content* of the prescription should be considered *irrelevant* for establishing the subject's obligation to obey. Nevertheless, Wolff argues, an

[16] Wolff, *In Defense of Anarchism*, p. 13.
[17] Wolff, *In Defense of Anarchism*, p. 40.
[18] Wolff, *In Defense of Anarchism*, p. 18.
[19] Wolff, *In Defense of Anarchism*, p. 6.

autonomous moral agent cannot commit itself to obeying the directives of authority even when they do not correspond to the agent's best judgment about what ought to be done. However, he acknowledges that there are cases when an individual preserves her moral autonomy while obeying a command irrespective of her own judgment concerning the goodness of what is commanded. In Wolff's example, a man who finds himself on a sinking ship where all passengers are obeying the captain orders' concerning the lifeboats might decide that, under the circumstances, compliance is the best course of action, since the confusion caused by him disobeying the captain would be generally harmful. What makes the man decide that compliance is desirable is not his judgment about the merits of the captain's orders, but rather the fact that these orders have been issued. Nevertheless, it should also be noticed that the *mere* issuance of the captain's orders is not sufficient to determine compliance. The decision to obey depends on whether certain *additional* conditions are met, e.g. that everybody else is obeying the captain, that failure to comply would be generally harmful. As long as the man judges for himself that these *additional* conditions are both decisive and fulfilled, his autonomy is preserved despite his decision to comply with the captain's order without judging their content.

The question is, of course, whether the analogy between the decision to comply with the commands issued by the captain of the sinking ship and the decision to accept political authority really holds. Broadly put, an individual might think that social security and order are values of ultimate importance. Moreover, he might think that these values could only be preserved as long as each member of society complies with the directives of that society's political authority. To the extent that the individual's obedience is not

unreflective, that is, to the extent that he himself judges that his compliance is desirable, it is not clear why his decision would amount to surrendering his autonomy.

Against the aforementioned analogy, it can be claimed that there are important differences between deciding to accept the orders of the captain of the sinking ship and deciding to accept the directives of authority. The fact that an individual decides to comply with certain commands without attempting to judge their content on a *particular occasion* makes it possible for him to preserve his autonomy, since he will be able to judge whether the mentioned 'additional conditions' obtain. By contrast, one's decision to comply with authoritative commands *in general*, which implies a *constant* willingness to do whatever the authority that issues them, seems to leave no room for one's autonomy. The right to rule claimed by political authorities' involves the right to pass laws that regulate almost every aspect of a society's life. This being the case, presumably it will be very difficult for an individual to judge in advance if certain additional conditions are likely to obtain or not. A way out of these difficulties is to reply that forming and passing judgments about whether the mentioned additional conditions obtain will be easier whenever individuals conceive of values such as social stability and order as impossible to be overridden. Also, as H. G. Frankfurt argues, it is most likely that the practice of holding a constitutional framework constant will increase individuals' reassurance that such things like 'the relevant additional conditions' will obtain.[20]

On the other hand, one could be willing to accept that the analogy between complying with orders on a sinking ship and accepting political authority holds and, at the same time, suggest that this analogy makes Wolff's position more plausible. On this view, what the analogy establishes, in fact, is that in order to follow authoritative

[20] H. Frankfurt, *The Anarchism of Robert Paul Wolff*, pp. 410–411.

commands, the mere fact that they have been issued by a certain authority is not sufficient. What is decisive is that *other* conditions obtain. Thus, in deciding whether to accept political authority, the individual actually assesses *independent moral reasons*. This would leave unaffected the standpoint of philosophical anarchism since, as Simmons argues, the anarchist could agree that although the law has no moral standing of its own, the conduct required by it is often morally obligatory due to such independent reasons (e.g. disobedience would frustrate reasonable expectations).[21]

Still, recall that the thrust of Wolff's argument consists in the claim that the conflict between preserving one's autonomy and accepting as binding the directives of authority is inherent, and unavoidable. However, if Wolff accepts both that the individual is allowed to subject himself to political authority (even if only because there are independent moral reasons that plead for obedience), and that the individual's autonomy is not precluded in this case, this will diminish the force of his argument for the existence of an inherent conflict between autonomy and authority. In this sense, the following reply to Wolff's argument from autonomy seems available: for individuals who accept the existence of moral reasons that tip the balance in favor of obeying the law, acting upon them by participating in the creation and preservation of political societies is, in fact, an aspect of the actual exercise of their autonomy. This type of autonomy, usually labeled as 'wide autonomy', is seen by some as conceptually incorporating in its very proper functioning the acknowledgement of the existence of a moral reason to obey the law. The difficulty with this position is that it operates a certain misconstruction of the philosophical anarchist's position, for, as it has been observed, in denying political

[21] A. J. Simmons, *Philosophical Anarchism*, pp. 13–17.

authority "the anarchist takes a narrower view of autonomy".[22] This narrower view is one that denies precisely the existence of such a reason for obeying the law.

Still, if we both take Wolff's concept of autonomy to be an example of the narrower view and keep to the character of a robust moral constraint that he gives to it, then it can be objected that autonomy excludes not only any kind of submission to authority, but also the moral appropriateness of any elements of precommitment to others that there can be, including morally innocent ones. Chiefly among excluded elements of precommitment will figure those that contribute render 'one-to-one' interpersonal relationships meaningful (the ability to enter promises and contracts *inter allia*), as well as those necessary for minimal participation in the social and/or political life of a collectivity.[23] The objection, however, seems to underscore an important element that underlies the 'standard' position of the philosophical anarchist, namely the insistence that individuals' decisions for actions have to be made considering only those reasons for actions that they *already accept*.

Later on, I will indirectly return to this objection and to other important related points, while discussing, with reference to Scanlon's seminal discussion of promises and of what is the most plausible interpretation of the requirement that making a promise must make a normative difference. For now, I must stress only that a discussion about the

[22] The observation is from Christopher McMahon's *Autonomy and Authority*, p. 304. Though relevant for its point, the observation does not seem to be an accurate analysis of all there is implied by Wolff's position.

[23] Referring specifically to Wolff, Keith Graham suggests that the most important resource for "coping with a sponsorship of autonomy so unqualified [as Wolff's], that threatens to engulf any thought of democratic commitment", lies in "the idea of collective identification" (p. 128, n 25). He then goes on to claim that "at the very least, collective identification is incompatible with retention of personal autonomy *on every occasion* when the collectivity acts" (Graham, *Practical Reasoning in a Social World*, p. 128, emphasis added). Graham's view is rooted in his previous analyses concluding that Wolff's conception is inconsistent even if we attempt to picture it as merely directed at minimizing individual heteronomy, rather than at making the exercise of autonomy an absolute duty. See Graham, *Democracy and the Autonomous Agent*, and *The Battle of Democracy: Conflict, Consensus, and the Individual* (sections on Wolff).

status of the second type of elements of precommitment that the objection mentions could benefit from noticing that it is ambiguous whether, for Wolff, the exercise of autonomy (as manifested by one's taking responsibility for one's actions) is an absolute duty impossible to override, or just one among several values to be collectively realized. One other stress on the ambiguity involved in Wolff's concept of autonomy is put forward by John Horton in his *Political Obligation*. In Horton's view Wolff does not seem to distinguish between the use of autonomy as a mere presupposition and its use for designating a moral ideal. Even if he would, the objection goes, he would anyway run into troubles. Also ambiguous are Wolff's treatment of the issue of conflict between autonomy and other human values and ideals (should we, under any circumstances, stick to the idea that it is immoral to make a compromise and restrict autonomy, in order to achieve an equilibrium, or stability, or personal and social peace, or some ideal other than autonomy?), as well as that of the issue of what are the exact contours of the duty obligation of autonomy (does it demand that each and every action we take should, in the final instance, make the object of independent individual rational assessment, or is this the case only in what the moral aspects of its application?).[24] In fact, there seems to be a connection either between all these ambiguities in Wolff, or at least between some of them. Barred an understanding of autonomy as an absolute duty, it seems that the argument which makes it inconsistent with the exercise of legitimate authority fails. This

[24] For passages in Wolff that seem to support both interpretations of this issue, see Wolff, pp. 12–13, and compare pp. 14–15. For the view that "whatever the scope of control over one's own life turns out to be, there is an incompleteness in the ideal of autonomy as articulated [here]", see Graham, *Practical Reasoning in a Social World*, p. 18, and his subsequent discussion. For general discussion of the (ambiguities involved in the) uses of the term 'autonomy', see Gerald Dworkin, *The Theory and Practice of Autonomy*, pp. 5–20, and *Autonomy*.

understanding, however, seems to be only available in the case of moral reasoning, and to remain problematic when considered within the broader realm of practical reasoning.[25]

2.3 The surrender of judgment thesis and the charge of irrationality

The previous remarks open up the possibility of looking at philosophical anarchism in the following, somehow indirect, way. To begin with, one can claim that people's voluntary cooperation for creating and continuing political societies seems to comprise, as a matter of fact, the undisputed belief in the existence of a moral reason to obey the law. In order to create a certain normative plausibility for this tenet, one can add the qualification that the invoked reason need not be seen as a non-derivative one. Rather, the reason can be seen as a consequence of following up on the constraints imposed upon the cooperating individuals by 'the division of moral labor', constraints in virtue of which government are allowed to take certain steps towards deciding how their subjects should act in certain situations. On this view, if all the subjects of a government would take government's decisions as reasons for action and this practice would be a matter of common knowledge, there would be no need for the use of coercion in order for that political society to exist. This not being the case, an intuitively plausible circumstance is clarified, namely that the constraints following from the division of moral labor will show up primarily in cases when acting upon the moral reason of obeying the law will require individuals to act not only against their best 'overall' rational deliberation, but against

[25] For similar points, see Graham's discussion in Graham, *Practical Reasoning in a Social World*, p. 16–20, and den Hartogh, *Mutual Expectations*, p. 128–129.

their best *moral* judgment of the case.[26] Moreover, remembering that the denial of the existence of a moral reason to obey the law is at the core of the anarchist's position, one is left with the main aim of showing, in order to answer the anarchist's challenge, how moral reasons other than this one can justify political obligations. The problem, of course, is that at least some of these remaining moral reasons are among the ones advanced by the philosophical anarchist in order to deny the existence of such obligations in the first place. This being the case, what one must carry out is the apparently paradoxical task of showing "how the reasons which support a judgment that the government is mistaken can nevertheless justify cooperating with it".[27]

For taking up this task, one must try to assess further the relationships that hold between autonomy, authority, obligation and (both individual and, as I shall argue, collective) rationality. Moreover, I believe that one must begin to do so with a necessary first move, namely with an attempt to formulate a strong version of the argument advanced by the philosophical anarchist against the existence of a moral obligation to obey the law. Discussing such a formulation would provide additional strength to any argument which is successful against it. For reasons already presented, I do not believe it is possible to employ Wolff's formulation. However, before attempting to put forward what I take to be the most powerful version of the argument for philosophical anarchism, let me note that something more seems to be at stake in this entire discussion, something that goes beyond merely casting doubts on the tenability of the anarchist position.

Both the need for further analysis of the relationships between autonomy, authority, and obligation, and the inadequacy of using Wolff's views as a paradigm for

[26] Such is the case, for instance, when one's obeying authoritative directives will imply cooperation to support the implementation of governmental measures and policies that one regards as morally mistaken.
[27] Christopher McMahon, *Autonomy and Authority*, p. 305.

philosophical anarchism have also been suggested by Leslie Green. Arguing against the possibility of a plausible deduction of the character of duty that Wolff attaches to autonomy, Green noticed that the thrust of Wolff's argument seems to be not on the (postulated) duty to be autonomous, but on the underlying conception of rationality. As Green writes:

> Reason itself seems to require that we always do what is best on the balance of reasons as we see them, whereas authority claims adherence contrary to the balance of reason and thus seemingly contrary to reason itself.[28]

According to this reading, the emphasis of Wolff's argument is not on the fact that an aspect of neglecting one's duty to observe moral standards and abide by them is always involved in the act of surrendering one's judgment, but on the charge of *irrationality* that becomes possible at this point. To illustrate his view, Green quotes David Gauthier, claiming that "an appeal to authority – to requirements imposed by authority – is an alternative to an appeal to reasons – to requirements based on reasons for acting".[29] This stance, rightfully criticized by Green, reduces the widest normative powers to be ever plausibly reclaimed by the requirements of authority to the status of advices. But something seems unclear here. In his critique, Green stresses the fact that Gauthier's suggestion, albeit intended to pinpoint a tension which "is part of the very concept of authority", ultimately confuses the very understanding of authority relations.[30] Still, while not halting in one version or another of the more common objections to the

[28] Leslie Green, *The Authority of the State*, pp. 25–26.
[29] The quote is from Gauthier's *Practical Reasoning*, p. 139.
[30] Leslie Green, *The Authority of the State* p. 26: "The oddity is that this does not accurately describe authority relations as understood by those who either accept or reject them". He goes on to notice that the distinction between authority and advice "is revealed both in the intentions of its subjects and in their different reactions to non-compliance" (ibid.)

'conceptual argument', Green proceeds by discussing the relationship between (expert) advice and authority.[31]

However pertinent, his reading seems, due to its taking this direction, somewhat unexpected: since he notices, importantly, that Wolff seems to make autonomy come in contradiction with authority by allowing the former to 'ally' itself with rationality, one would expect his critique to concentrate in at least the same measure on the soundness of Wolff's views about this important relationship.[32] Partly as a response to this felt need, I will later on look at the way in which the relationship between autonomy and rationality is relevant for an individual's reasoned decision to cooperate with political authority (or refrain from cooperating with it).

Before that, however, one more issue stands in need of clarification: in insisting that Wolff's argument makes room for the charge that it is always irrational to surrender one's judgment, the proponent of the rationality reading too hastily abandons Wolff's main claim, namely that it is *immoral* to ever submit to the requirements of authority. I believe, however, that this serious objection can be easily dodged if one succeeds in showing that, in fact, the philosophical anarchist can attempt to maintain the coherence of his position even while dropping this claim. That Wolff's claim is not necessary for the argument in favor of philosophical anarchism seems, at first glance, good news. In fact, a strong argument along these lines does exist, and it is worth analyzing. As it will turn out, the argument fails to support philosophical anarchism, but its failure opens up a promising way to proceed.

[31] Leslie Green, *The Authority of the State*, pp. 26–29.
[32] Here I do not mean to criticize Green for 'incomplete treatment of the relation between autonomy and rationality' (in fact, he does consider it later on). I only try to make a point about an off beam turn of his argument, and to justify the direction of my own argument. For a point that Green does seem to overlook, by going directly for the 'rationality' reading of Wolff, see below.

In his *Mutual Expectations*, Govert den Hartogh puts forward what I take to be a much needed plausible argument that a philosophical anarchist of Wolffean descent would find attractive. Let me briefly reconstruct his argument.

1. A's decision, made after careful consideration of all available reasons, that he has, all things considered, a reason to φ, together with the absence of any reason for which A should distrust his judgment on the matter, make it irrational for A not to φ.
2. It is an essential feature of a political unit (a government, a state, or any similar organization that could be properly called political), that it claims authority over the actions (and not only, nor even necessarily over the beliefs) of its subjects. Arguably, this is a claim to supreme (i.e. unlimited) authority.[33]
3. To the extent that an agent does indeed exercise such an authority over A, the sheer fact of that agent's requiring A to do something constitutes a decisive reason in favor of A's doing it. This holds irrespective of the precise content of the authoritative agent's requirement, with some exceptions (to be imagined along the lines envisaged by Hart when claiming that "grossly immoral promises do not bind").
4. Suppose now that the claim advanced by the political unit is legitimate.
5. From (3) it results that occasions exist when the fact that a political unit requires A not to φ makes it rational for A not to φ, even if the balance of reasons indicates that A has reason to φ and no reason to mistrust his judgment on the matter at

[33] Such is, at any rate, the claim advanced by law, and it's one which is directly relevant for the issue of political obligation. See Raz, *The Morality of Freedom*, p. 76–77, and Raz, *The Obligation to Obey the Law*, pp. 115–120.

stake. In other words, considerations of political authority override, on occasion, all things considered individual considerations about the balance of available reasons.

6. (5) contradicts (1). Hence, (4) is false: the claim to authority essentially advanced by political units is invalid and, therefore, no political unit is ever legitimate.

Despite obvious similarities, the previous argument is not identical with Wolff's. For one thing, it surely does not lend support to his claim that accepting authority requirements is immoral. For another, it does not purport to prove the same conclusions. As I pointed out, Wolff wants to be able to claim that we are bound to always form and pass judgments on the balance of reasons as they apply to us; and, moreover, that we are also bound to always follow up on these judgments, by acting in accordance with them, regardless of the consequences or the success of such actions.[34] This move, while designed to increase the overall coherency of his project, ends up undermining it. By contrast, what is required for the anarchist argument (in the version put forward above) to go through, is that we accept the weaker, conditional claim that if one manages to form a sound view on a given matter, eliminating all reasons to mistrust one's own judgment (and, hence, going beyond the possibility of rational malfunction), it would be irrational not to act on it. Wolff's strong claim is unnecessary.[35] In fact, as den Hartogh points out, claim (1) is not only true, but a truism. This is emphasized by the fact that its negation – which states that A might have reasons to φ, i.e. that it might be rational for A to φ, even though it is not irrational for A not to φ – is clearly paradoxical.

[34] He does accept, however, the epistemic authority of the medical expert. See Wolff, p. 15. See also my discussion of the case of practical epistemic authority, below.
[35] Den Hartogh, *Mutual Expectations*, p. 130.

One interesting way of bringing out the difference between Wolff's claim and (1) is to note that only the former can be denied by employing Raz's useful distinction between two different understandings of reasons for action. According to Raz, reasons for action can be understood either as reasons for conformity, or as reasons for compliance. While people conform with a reason for a certain act if they perform that act in the circumstance in which that reason is a reason for its performance, compliance with a reason involves something more, namely, a display of appropriate sensitivity to (the importance of acting on) that reason.[36] More concretely, A conforms with a reason to φ simply if A φ-es, without necessarily being aware of there being a reason to φ. By contrast, A complies with a reason to φ if A realizes that she has a reason to φ and consequently acts on that reason. If Raz is right in claiming that only conformity is essential to acting for a reason, then Wolff's stronger claim to the effect that (i) we are bound to always form and pass judgments on the balance of reasons as they apply to us, and that (ii) we are also bound to always follow up on these judgments, by acting in accordance with them, regardless of the consequences or the success of such actions, is false.

Before moving on, it is important to note that, as presented, the argument is flawed, in spite of claim (1) being true. As den Hartogh rightly points out, the flaw is to be found in step (5), for while authority supplies content-independent reasons, it does not necessarily require surrender of judgment. The mere fact that a reason is content-independent does not entail, as a matter of conceptual necessity, that the reason in question is all things considered overriding. There is no contradiction in claiming that a content-independent reason is only one of the reasons to be weighed in the overall

[36] Raz, *Practical Reasons and Norms*, pp. 178–181.

balance of reasons, and can thus be overridden. It seems to follow, then, that political authority properly understood cannot, on pain of irrationality, require surrender of judgment.

This point relates to Raz's notion of exclusionary reasons, as will become clear in the discussion to follow. The very existence of exclusionary reasons seems to undermine the conclusion of the last paragraph, according to which legitimate authority cannot require one to disregard one's *reliable* judgment of the balance of reasons. However, it is worth stressing that, should exclusionary reasons be able to undermine the reliability of one's judgment of the balance of reasons, the reasons they purport to exclude would have to have been confidently judged to be valid. But, as den Hartogh argues, no valid exclusionary reasons exist which exclude reasons of this type. A promising way to proceed in attempting to see why this is so is to notice first that Raz puts together two classes of exclusionary reasons, namely those provided by epistemic and by coordinating authority. He does so in order to delimitate his own position from the subjective conception of practical reasons.

On his view, reasons are given by facts, not beliefs about facts. One's having a mistaken belief about a fact may lead one to mistakenly believe that one has a reason, which is why often talk of reasons is couched in terms of beliefs rather than facts. However, even in such cases, it is a certain state of the world (mistakenly appraised as it might be) and not the belief about it that constitutes a reason to act. Raz goes even further than this by suggesting that in practical deliberation, i.e. in the process of weighing reasons against each other, belief should be allowed no substantial role. His opinion on

the matter is, in fact, a reply to Stephen Perry, who notices that Raz's view seems to run into trouble:

> The idea […] presupposes a generalization of the subjective conception of an exclusionary reason. According to this generalization, a subjective second-order reason is a reason to treat a reason as having a greater or lesser weight than the agent would otherwise judge it to possess in his or her subjective determination of what the objective balance of reasons requires. (An exclusionary reason is then just the special case of a reason to treat a reason as having zero weight.) Second-order reasons as thus defined will be referred to as reweighting reasons. Notice that the idea of a reweighting reason only seems to make sense if it is regarded as a possible strategy upon which agents might rely in their subjective practical determinations about what ought to be done […]. *The idea that reweighting could take place at the level of the objective balance of reasons does not even seem to be coherent, so that a generalization of the sort being discussed is possible with respect only to the subjective and not the objective conception of an exclusionary reason*.[37]

In reply, Raz claims that Perry's notion of a weighting reason can be brought into line with his earlier argument to the effect that reasons for action are facts rather than beliefs about them, by taking Parry's notion of a weighting reason to mean the following: "Certain facts are reasons for assigning other facts, which are reasons for action, a greater or lesser weight than they would otherwise merit."[38]

At this point, the following objection can be mounted. Raz's insistence in claiming that only facts, and not beliefs, can justifiably figure in the process of weighing reasons against each other seems to obscure the fact that, when belief is less than certain, it would be irrational not include at least a judgment about the *probability* of this belief's being true in the process of balancing reasons. As den Hartogh argues:

> If I have reason to mistrust my own judgment to a certain extent, then my assessment of the reliability of my judgment (and my assessment of the

[37] Stephen R. Perry *Second-Order Reasons, Uncertainty and Legal Theory*, pp. 932–933 (emphasis added).
[38] Raz, *Facing Up: A Reply*, p. 1178, (emphasis added).

> reliability of my assessment) should be taken into account in the weight I attribute to any purported reason. But then, it seems, what I am weighing in the balance of reasons after all, is my belief about the [fact], for the weight of my reason is determined by the probability I believe my belief to have of being true.[39]

However, this point about probability is moot, since it is by no means clear whether probability itself should be analyzed in belief-dependent, rather than in belief-independent terms, as den Hartogh would seem to imply.[40] Still, there are other elements of a clearly subjective nature which must be factored in a proper understating of the process of weighing reasons. Such are the epistemic limitations of deliberating agents (e.g. lack of expertise, time constraints on decision-making etc). Whatever the epistemic status of 'the probability factor', it is obvious that the epistemic limitations end up by introducing a subjective element in the agent's weighing of reasons. I do not mean to suggest that striking a balance of reasons has a preeminently subjective character. Rather, my move is meant as an attempt to emphasize the necessity of taking the agent's point of view into account. This, I believe, is consistent with Raz's claim that reasons, as provided by facts, are objective (i.e. belief-independent). There is no contradiction in claiming that reasons are objective and, at the same time, insisting that our reasoning capacities must have a minimal 'filtering' role in the reasons-guided process of decision-making. More concretely, the point is that, even if reasons are given by objective facts, it is only by assuming flawless epistemic abilities that one can claim to have immediate access to this realm of objective reasons. Yet, since we are obviously not perfect knowers, this assumption would be unreasonably strong. What follows is that we must accept that the agent's access to the reasons provided by the relevant facts is always limited by the

[39] Govert den Hartogh, *Mutual Expectations*, p. 133.
[40] Govert den Hartogh, *Mutual Expectations*, pp. 133–134.

agent's imperfect epistemic capacities. Still, these remarks need not end up in an outright denial of the objectivist stance on reasons for action, for they do not entail that it is permissible for the agent to knowingly neglect a reason he takes to be valid. In other words, the mere fact that our imperfect epistemic capacities are factored in the general process of practical deliberation should not be seen as a concession to skepticism. Although we are aware of the fact that we are not ideally placed from an epistemic point of view, this does not entitle us to question every possible judgment regarding reasons for action. In some cases, we simply know that the judgment we came up with is the best judgment available to us.

Chapter 3
Models of political authority

In this chapter, I place under critical scrutiny Joseph Raz's model of authority, which is widely regarded as the most plausible model for the justification of political authority available, as well as argue for the need for adopting a model of collective rationality, a need resulted both from the necessity of finding a mediating factor in the announced conflict between authority and autonomy, and from the inability of Raz's model to justify political obligation. A scrutiny of the proposed framework is put forward in the last section of chapter 3, in an attempt to outline one of the important directions of the second part of the thesis.

3.1 The conceptual argument

Let me start the discussion of authority by saying a few things about what is nowadays called *the conceptual argument* for political authority and political obligation.[41] The starting point that made authors of different orientations advance the conceptual argument is the observation that, in many different uses of the topic of political obligation, a peculiar aspect of *the very concept* of legitimate political authority is that it indicates the existence of a moral duty to obey the laws issued by the authority. The underlying idea is that one cannot both acknowledge that a government is legitimate and, at the same time, deny that one has a duty to obey. Hannah Pitkin, for instance, writes

[41] This label was coined by Carole Pateman, in her *Political Obligation and Conceptual Analysis*.

that "part of what 'authority' means is that those subject to it are obligated to obey".[42] Richard Flathman believes that the one who does not understand the conceptual link between political authority and the duty to obey simply does not understand the semantic rules guiding the use of such concepts as 'authority' and 'law'.[43] In a different note from his own earlier writings, Joseph Raz also advocates for a version of the conceptual argument when writing "I do not wish to obscure the fact that exercise of authority involves a claim that those subject to it have a moral duty to obey it".[44] The most important thing is, of course, to decide what can we coherently make out of this thesis. An available option is to interpret the aforementioned argument as merely stating that the *definition* of such terms as 'political authority' and 'legitimate government' include, of a logical or analytical manner, a duty to obey. What is argued is that there is a conceptual confusion in supposing that the former concepts could exist without the latter, and "since this is just what those who seek a general justification for political obligation do suppose, their project is fatally undermined", an independent justification of the obligation to obey being "neither necessary nor possible."[45] Apart from constituting a strange exercise in terminological definition, this interpretation offers no interesting insight – it does not advance the debate much. As it has been noted, the conceptual argument seems to "explain away the problem [of political obligation]".[46] It is most likely that anyone seriously bothered by the many diverse cases in which the alleged obligation of

[42] Hanna Pitkin, *Obligation and Consent* – II, p. 60.
[43] Richard Flathman, *Political Obligation*, pp. 89–90.
[44] Raz, *Government by Consent*, p. 77; but compare Raz, *The Obligation to Obey the Law*, p. 233, where he argues that "there is no obligation to obey the law", and suggests that "there is not even a prima facie obligation to obey it".
[45] John Horton, *Political Obligation*, p. 138.
[46] Horton, *Political Obligation*, p. 138.

obedience seems to be in striking conflict with the demands of individual morality will have to refute this strong version of the argument.

It then seems that a proper insight into the issues of political authority and political obligation would have to advance the claim that, rather than being a merely conceptual one, it is an *independent*, *substantial* feature of political authority justifies individuals' having a duty to obey. The thesis that 'any government that is morally justified in coercing its citizens has a right to be obeyed' might seem to constitute a plausible example of such a feature: taken together with the claim that the legitimate political authority's right to rule is the logical correlate of an obligation to obey, this thesis would be able to set a standard for circumscribing the obligation in question. Still, the problem that one has to deal with in order to successfully assess the merits of this thesis is that of looking at whether it can be accepted that a legitimate state would have the right to coerce its subjects. This claim could be considered, following the Weberian line of thought, as being in itself part of a somehow different conceptual argument – one for establishing the monopoly of force that a legitimate state should have. Indeed, since Max Weber's work, it has been a common tendency among political theorists to think of political societies as being forms of social life in which governments posses a monopoly on the legitimate use of coercion.

However, arguments placing an exclusive focus on governments' monopoly of coercion have been criticized, rightfully I think, on several grounds. Herbert Hart for instance, rightly pointed out that a government's actual attempt to coerce an entire population would give rise to insurmountable empirical difficulties.[47] This practical impossibility seems to leave room for different ways of looking at why people establish

[47] Hart, *The Concept of Law*, pp. 21–22.

social forms as complex as political societies are. One such way, following directly from Hart's observation, is to advance the equally empirical claim that people voluntarily cooperate in order to create and maintain political societies. While acknowledging that some people only cooperate because they are coerced, theorists who use this argument claim that it is empirically provable that most of them do so voluntarily. Another way has a different character, but the same substance: it advances the normative claim that, since establishing political societies requires a collective effort, individuals should be seen as voluntarily cooperating towards their establishment, either out of habit or because they judge that there are good reasons for so doing. In his *The Principle of Fairness and Political Obligation*, George Klosko starts by observing that "the existence of strong general feelings that we have political obligations…is supported by our most basic feelings about politics", and goes on to claim that it is "obviously true that most people believe that they have obligations to their governments".[48] Neither one of these two ways of dealing with the difficulties involved in conceiving of governments as possessing a monopoly on the legitimate use of coercion attempts at denying the special place that governments occupy within political societies. Rather, the proposed view is one in which governments occupy a position of authority: they issue binding authoritative directives, and their subjects comply with them. It is the view that individuals voluntarily obey the laws and commands of the government.

[48] Klosko, *The Principle of Fairness and Political Obligation*, p. 22. Later on, he also claims that "As a corollary of this belief [that they have obligations to their governments], most individuals believe that their governments are legitimate and, by implication, acceptably fair" (p. 68).

3.2 The Razian model: a reconstruction

Joseph Raz spells out a general justificatory model for practical authority, and attempts to show further how this functions within a plausible model of political authority. According to Raz, many of the problems posed by authority in general are created by the employment of an oversimplified view about the requirements of practical reasoning, and by our consequent failure to recognize that authority is a *practical* concept. The origins of Raz's views are to be found in his debate with legal philosopher H. L. A. Hart on the issue of the separation between law and morality.[49] In his *Essays on Bentham*, Hart claimed to find "little reason to accept…a cognitive interpretation of legal duty in terms of objective reasons or the identity of meaning of 'obligation' in legal and moral contexts which this would secure". He went on to explain that:

> […] to say that an individual has a legal obligation to act in a certain way is to say that such action may be properly demanded or extracted from him according to legal rules or principles regulating such demands.[50]

In reply, Raz observes that, since statements asserting that each person ought to do **X** are logically equivalent to statements asserting that each person has reasons to do **X**, such statements must represent either moral considerations or considerations of each addressee's interests. However, since person **A** cannot justifiably tell person **B** to have a certain conduct only because it is in person's **A** interest, it follows that, in order to grant her claim a normative status, person **A** must claim that person **B** has a moral

[49] For a good discussion of the Hart vs. Raz debate, and of most of the arguments that I present below, see Matthew H. Kramer's *Requirements, Reasons and Raz: Legal Positivism and Legal Duties* (esp. pp. 378–381).
[50] Hart, *Essays on Bentham*, pp. 159–160.

responsibility to promote her (i.e. person **A**'s) interests. This is the claim that authoritative practical claims are interest-independent. In addition to this claim, Raz reminds us that practical authority's directives concerning legal and political duties should be seen as claiming 'authority over conduct' (i.e. that they are directives about how individuals should behave), and that they consequently assert that individuals have reasons to behave in the prescribed way. From this two premise-claims, Raz argues, it follows that the reasons for action included in a practical authority's directives must be moral reasons. Raz concludes that:

> No system is a system of law unless it includes a claim of legitimacy, of moral authority. That means that it claims that legal requirements are morally binding, that is that legal obligations are real (moral) obligations, arising out of the law.[51]

It becomes obvious that the force of Raz's argument greatly depends on the conception of moral reasons that it employs.[52] Raz argues that authoritative directives possess the feature of offering *exclusionary, content-independent* reasons for action, and that the subjects of a given *de jure* authority should obey its directives "even if their personal belief is that the balance of reasons on the merits is against performing the required act".[53] He points out that in many cases of practical decisions our reasons for undertaking a certain course of action are structured on different levels. In his words, in many such cases we appeal not only to first-order reasons but also to second-order reasons. First-order reasons are ordinary reasons for action such as desires, interests,

[51] Raz, *Hart on Moral Rights and Legal Duties*, p. 131.
[52] What also becomes clear at this point but will become important only later in the text is the fact that the normativity that Raz ascribes to legal norms is an *internal* one.
[53] Raz, *Authority and Justification*, p. 120.

needs, etc. They always claim that one should act in conformity with them. Conflicts between first-order reasons are always resolved by comparing and balancing their relative weights: one such reason will be defeated by another reason of the same level only if it is going to be outweighed by it. Second-order reasons are the reasons that we recognize either for or against acting on first-order reasons. In the first case, they are positive second-order reasons; in the second case, they are *exclusionary* reasons. Conflicts between a negative second-order reason and a first-order one are not solved as a result of the former's reason outweighing the latter; rather, the former reason excludes the possibility of acting by resting solely on the balancing between the latter reason and other first-order reasons. Precisely in this sense second-order reasons are *exclusionary*: they exclude other reasons by being different *in kind* and not merely by possessing a different strength. The binding force of exclusionary reasons is, according to Raz, given by their being both categorical and *prima facie*: they are categorical in that they exclude and not merely outweigh reasons for not performing a certain action, and they are *prima facie* in that they may not exclude *all* contrary reasons.

This is, however, only the first part of the story. There is a second essential feature of the Razian view of what is implied in the notion of reason for action offered by the directives of a practical authority, namely that of content-independence. Content-independent reasons are inherent to the very fact that one has issued a certain demand. That the reasons provided by practical authority exhibit the feature of being content-independent becomes clear if we understand that such reasons are somehow external to the very actions that they are reasons for. As Raz writes,

> A reason is content-independent if there is no direct connection between the reason and the action for which it is a reason. The reason is in the apparently 'extraneous' fact that someone…has said so, and within certain limits, his saying so would be reason for any number of actions, including (in typical cases) for contradictory ones.[54]

Taken together, the content-independent and the exclusionary character of certain reasons account for the fact that, whenever such a reason is presented, one is seen as having both a new reason to act as required *and* a reason to forbear from acting on the balance of those reasons one formerly had, even if this balance would have indicated a course of action different from (or even contrary to) the one indicated by the new reason. Raz calls such reasons 'protected reasons', and argues that it is in this way that we should understand the functioning of most binding commitments and, in particular, that of authoritative directives.[55] Such commitments are cases in which a person has a reason for performing a certain action *and* an exclusionary reason not to act on some of the reasons for not performing that action. To take an example, a vendor's announcement that the shop is closed excludes any considerations about the merits (e.g. about the urgency) of one's need for a certain supply: the vendor does not take the view that the desirability of closing the shop outweighs that of staying open for some more time, she simply refuses to consider the issue any further. For illustrating his conception of authority Raz however uses the example of two litigants who refer their dispute to an arbitrator. Whichever it is, the arbitrator's decision is meant to be a reason for action for the two litigants: they should do as the arbitrator says because he says so. The arbitrator's decision has two important features. The first one that it is, as Raz says, "meant to be based on the other reasons [of the litigants that are relevant for the case], to sum them up and to reflect their

[54] Raz, *The Morality of Freedom*, p. 35.
[55] Raz, *The Morality of Freedom*, p. 18.

outcome". In this sense, the decision is a 'dependent reason' for the litigants. The second feature of the arbitrator's decision is that it is a reason which replaces the prior, conflicting reasons of the litigants. In this sense, Raz calls it a 'pre-emptive reason'. The whole point of discussing the case is to emphasize that, once they agree to hand over the evaluation of their prior reasons to the arbitrator, the litigants are "excluded from later relying on them" and, thus, their initial reasons "cannot be relied upon once the decision is given."[56]

The relevant conclusions drawn from this example are generalized in three theses the combined work of which is, according to Raz, characteristic of any *de jure* practical authority. The first one is *the dependence thesis*, "a moral thesis about the way authorities should use their powers".[57] It states that:

> […] all authoritative directives should be based, in the main, on reasons which already independently apply to the subjects of the directives and are relevant to their action in the circumstances covered by the directive.[58]

The second one is *the pre-emption thesis*, which states that an authority's requiring the performance of an action constitutes a reason for that action's performance which "is not to be added to all other relevant reasons when assessing what to do, but should replace some of them".[59] Whenever the conditions imposed by the dependence and the preemption theses are met, there will be only one "normal and primary way of justifying the legitimacy of an authority". It will be one aimed at showing that the authority in question is more likely to act successfully on the reasons which apply to its

[56] Raz, *Authority, Law and Morality*, pp. 212–213.
[57] Raz, *Authority and Justification*, p. 129.
[58] Raz, *Authority and Justification*, p. 125.
[59] Raz, *Authority, Law and Morality*, pp. 214.

subjects, and that its acceptance will provide individuals with "a more reliable and successful guide to right reason".[60] This is expressed in the third thesis advanced by Raz, namely *the normal justification thesis*, which states that:

> The normal and primary way to establish that a person should be acknowledged to have authority over another person involves showing that the alleged subject is likely to better comply with reasons which apply to him (other than the alleged authoritative directives) if he accepts the directives of the alleged authority as authoritatively binding and tries to follow them, rather than by trying to follow the reasons which apply to him directly.[61]

Together, the first and the third theses constitute what Raz calls the service conception of authority. It is a view on which authority is seen as "mediating between people and the right reasons that apply to them".[62] Authority's main function is, on this view, to serve its subjects, and it purportedly does so by providing them with the possibility to advance their aims better than they could do on their own. The most important advantage of accepting Raz's model seems to be that of accounting for the claim that authoritative directives have a categorical normative force that does not depend on their weight: although not all of them have an equal force, they still seem to be categorical in a way in which considerations of self-interest, for example, are not. Moreover, due to the exclusionary character that the former considerations possess while the latter do not, this claim holds true even without implying the undesirable claim that every authoritative directive is weightier that every consideration of self-interest.

[60] Raz, *Authority and Justification*, p. 135.
[61] Raz, *Authority, Law and Morality*, p. 214.
[62] Raz, *Authority, Law and Morality*, p. 214.

3.3 The Razian model: an appraisal

All these matters constitute the backbone of Raz's overall argument for the justification of practical authority. However, an important problem arises. Since, due to the work of the preemptive thesis, accepting this model will always involve individual's "giving up one's right to act on one's judgment on the balance of reasons", it follows that, for an authority to be legitimate, a further requirement must be satisfied, namely that the exclusionary reasons presented by its directives be, at the same time, *valid* reasons. Therefore, although the momentary conclusion is that, in cases of conflict between the displaced first-order reasons and the secondary-reason that displaces them, if the presented exclusionary reason is valid then it will always prevail, the important question to be answered is 'what exactly renders such a preemptive reason valid?' Attempting to answer this question requires the development of two other aspects of Raz's arguments. The first one consists in a short discussion of some of his views about the characteristics of legal norms. The second one, closely connected with the first, explains further his characterization of practical authority.

Let me say a few things about why it is necessary to discuss the first of the two announced developments. In virtually every overview of the literature on political obligation, Raz is put in the camp of those contemporary authors who deny the existence of a general obligation to obey the law. At a first glance, this might seem to be both inconsistent with his argument that political authority involves the possession of a right to rule and, consequently, the existence of an obligation to obey the law, and straightforwardly puzzling vis-à-vis his entire enterprise of elaborating a complex construction of what political authority is about. Moreover, since Raz stresses the fact

that "the normal exercise of political authority is by the making of laws and legally binding rules", and that this is the reason "why much of the discussion of the justification of political authority is undertaken by a consideration of the obligation to obey the law", denying the existence of a general obligation to obey the law seems to endanger the very possibility that legitimate political authorities exist.[63]

In order to clarify this position one can start by observing, with Christopher Morris, that part of the problem follows from the fact that "the authority claimed by legal systems and states is especially hard to justify, even for reasonably just systems, as it is very extensive".[64] In Raz's terms from Practical Reason and Norms, this remark can be said to correspond to the claim that the authority of legal systems is both comprehensive (i.e. "[legal systems] claim to regulate any type of behavior") and supreme (i.e. "every legal system claims authority to regulate the setting up and application of other institutionalized systems by its subject-community").[65] It is due to this claim that individuals' obligation to obey the law is commonly seen as a general one, i.e. as holding "for all subjects, for all laws, on all occasions to which they apply".[66] Raz, however, argues that, regardless of how just a state will be, there will always be some laws that will not provide some subjects, on some occasions, with reasons for action. And it is precisely holding this position that leads him to deny that such a general obligation to obey the law exists. For understanding his views on these points it is necessary to clarify what is behind Raz's belief that a legal system claims both comprehensiveness and supremacy,

[63] Raz, *The Morality of Freedom*, p. 100.
[64] Christopher W. Morris, *Well-Being, Reasons, and the Politics of Law*, p. 819.
[65] Raz, *Practical Reason and Norms*, pp. 150–152. To these he adds a third characteristic of legal systems, namely that of being open.
[66] Morris, *Well-Being, Reasons, and the Politics of Law*, p. 819.

and this can be easily done by looking at this claim in connection with the normative status that Raz ascribes to legal norms.

Let me take a step back for a moment and mention an important criticism addressed to Raz's arguments. It is a two-horned criticism stating, on the one hand, that there is something wrong with the very concept of 'exclusionary reason' and, on the other hand, that even if this concept is sound, it cannot be convincingly argued, as Raz wishes, that this type of reasons is characteristic of legal norms. In reply to this criticism Raz argues that two distinct directions should be followed in order for his entire argument to be safe. The first direction is to try to justify the very concept in cause, while the second one is to try to show in what way does any proper characterization of a legal norm need to employ the concept. Note, however, that denying the success of the first direction will entail denying the possibility of success for the second, while the reverse is not the case. In other words, one can accept that the special kind of reasons that Raz labels as 'exclusionary' do exist, while negating that the characterization of legal norms should, as a matter of necessity, employ this concept. For increasing the fluency of the discussion and for better stressing its main point, let me start by briefly discussing the second of the directions that Raz suggest, direction which I find the most problematic.

As I already mentioned, Raz claims that legal norms are characteristically exclusionary reasons. Two observations are in place about this claim. The first one is that this is only so from the point of view of their (content-independent) nature, while from the point of view of their content they function as ordinary first-order reasons. The second observation is that, although legal norms are reasons for action, Raz argues that they do *not* possess a moral quality of their own. Rather than springing from their moral

character (i.e. from their importance as moral reasons), the exclusionary force of legal norms follows precisely from their being *legal* norms. Raz argues that understanding what law is must be preceded by understanding what law claims to be. To understand this (i.e. what law claims to be), we must recall two of Raz's remarks from *Authority, Law and Morality*. The first one is that a *de facto* authority "either claims to be legitimate or it is believed to be so".[67] The second remark is that "every legal system which is in force anywhere has *de facto* authority".[68] So doing will hence make us see that "the claim to authority is part of the nature of law".[69]Also, we should remember that, according to Raz, there are two ways of arguing for the normativity of a legal norm. One way is to argue that they are valid reasons; the other way is to argue that those affected by them *believe* that they are valid reasons. Consequently, the fact that legal norms are not valid reasons does not *necessarily* prevent them from being normatively important, although it may occasionally do so.[70]

However, the conclusions that one can justifiably draw from this conception about the normativity of legal norms essentially depend on what one means by "ways of arguing for the normativity of a legal norm". If by this label one understands that legal norms are capable of motivating action (when they are recognized as or believed to be valid), then one will have to admit that this normativity is not, in fact, a characteristic of legal norms, but rather of the individuals' attitudes (of recognition, belief, acceptance, etc.). If, on the other hand, normativity is to remain a characteristic of legal norms, then it seems that Raz must argue that legal norms become justified by the mere fact that those

[67] Raz, *Authority, Law and Morality*, p. 211.
[68] Raz, *Authority, Law and Morality*, p. 215.
[69] Raz, *Authority, Law and Morality*, p. 215.
[70] For both these points, see Raz's *Practical Reasons and Norms*, p. 170.

who are their subjects believe them to be justified and accept them. It is clear that Raz will be able to say that the law manifests the internal normativity that he initially ascribed to it only if he accepts the second of these interpretations. However, the only thing that follows from the fact that a legal norm is accepted or believed to be valid is the fact of *explaining* what are the reasons to think that the norm should be followed, and not the fact that the norm is, in itself, valid or justified in any other relevant way. To deny this would be to deny that there is a difference between having a reason for action and the mere belief in having a reason for action, and I do not believe that Raz would want to engage in this argument. He would most likely want to see the very *existence* of a reason for action as objectively assertible, at least for avoiding the catastrophic consequence (entailed by the possibility of its being subjective), that anyone who believes they have a reason to φ thereby have a reason to φ.

It then seems that Raz's argument fails to assess both the internal normativity of law that he initially envisages and the characterization of legal norms as being essentially exclusionary reasons. However, as I suggested before, denying that legal norms have this character does not by itself entail that 'exclusionary reason' is a vacuous concept. In order to what can be said about this issue, let me now turn to the first of the two directions suggested by Raz as a defense of his claims.

I will have little more to say at this point about the theoretical status of the concept of exclusionary reason. Still, I must note is that it is a misunderstanding to try to object its soundness by arguing from the premises a) that exclusionary reasons must necessarily be *valid* reasons and b) that Raz does not offer an all-encompassing account of the circumstances in which they become valid by really excluding first-order reasons.

This type of argument seems to overlook that all that is necessary for Raz's 'exclusionary reason' to constitute a sound concept is to point to the existence of at least *some* such reasons. It is obvious, as even the aforementioned example of the postal clerk seems to show, that there are circumstances that render valid reasons that seem to satisfy Raz's definition of an exclusionary reason. However, an argument of the same type can be couched in such a manner as to become relevant as a critique of Raz. It is important to mention it here, since later in the text I will continue to refer to and make use of the notion of 'exclusionary reasons'.

Instead of focusing on the question whether exclusionary reasons should be granted an independent conceptual status or not, the restated argument proceeds by focusing on the way in which they fulfill the role that they are supposed to, within Raz's argument for authority and in general. Steven Lukes, for instance, makes the point that there is an important range of cases and (public) areas in which the relationship between the directives of authority and individuals' reasons that they seek to displace is such that the displacement will not go according to Raz's scheme. According to Lukes, such a case is the one where the relationship between authority and reason is an intrinsic one, i.e. "where the objectives authority serves are internal to, that is shaped and sustained by, the authority relation itself".[71]. Innumerable examples seem to fit here, but the following can be thought of as exemplary. Within (the entire field of) religion, the authority of priesthood is not given by the attempt to *serve* those who place themselves under the authority of the priest, but by the attempt to further the truth expressed by religious dogmas. The charismatic leader who, as Lukes recalls following Max Weber, usually defines her follower's goals and whose legitimacy springs from her follower's belief in

[71] Steven Lukes, *Perspectives on Authority*, p. 213.

the extraordinary character/qualities of the leader in case. The case of a political party's action of prescribing the primacy of certain goals over others. Finally, the case of a parent's modeling of the self-understanding of her infant children and having the legitimacy to do so conferred exclusively by the role (of parent) that she fulfills. In all of these cases, as well as in several others, it seems clear that an authentic (i.e. *de jure*) authority is involved. Therefore, if Raz's approach is to live up to his stated goal of offering a "conception of the nature of practical authority"[72], then it should be able to account for these kinds of equally practical authority. As Lukes notes, the only possibility of doing so would be to deny that an authority can always claim legitimacy, and to maintain that "only if putative authorities guide their subjects extrinsically to 'right reasons' can their claims [to legitimacy] be justified"[73]. However, this boils down to the claim that the concept of a 'de jure authority that does not involve a claim of obedience' might be a sound concept after all, and this contradicts Raz's claim that authority always "either claims to be legitimate or it is believed to be so, and is effective in imposing its will on many over whom it claims authority".[74]

I do not know how Lukes' objection can be answered while at the same time keeping up with Raz's aim of providing a solid model for legitimate practical authority. On the contrary, I agree with his main contention that one can think of instances of practical authority that do not fit naturally into Raz's model, but rather have to be forced in it, as if placed on the Procrustean bed. Also, in what the dynamics of individuals' decisions about the weight they give to the directives presented by various practical authorities is concerned, I fear that it is only too seldom that things go according to Raz's

[72] Raz, *Authority and Justification*, p. 115.
[73] Lukes, *Perspective on Authority*, p. 214.
[74] Raz, *Authority, Law and Morality*, p. 211.

model. We rarely see people unfolding the reasons presented by a certain authority, making sure that the authoritative directives are based on reasons that apply independently to their actions, and only then allowing these directives to preempt the reasons for action that apply to them directly. Rather, people usually attach many other nuances to authority relations, and these go from trust, loyalty or utilitarian calculus to apathy, guilt or fear of social isolation. Moreover, when this is the case, it is rarely also the case that the relation of authority that should be in place is diminished, distorted or in other significant way impoverished by the presence of the mentioned nuances.

All these being said, instead of going for a rushed rejection of the Razian model of authority, we would be better off trying to preserve it in its main articulations, while reducing the scope of its application. In other words, we can reply to the former type of objection by saying that, instead of being put forward as an undifferentiated model for practical authorities in general, Raz's model is primarily concerned with political authorities. This possibility is considered by Raz as well, but he nevertheless prefers to "make no attempt to characterize the special features of those (i.e. of political authorities), as opposed to practical authorities in general."[75] In what the present text is concerned and due to its main preoccupation with finding a plausible answer to the challenge that the philosophical anarchist, placing the Razian model under the aforementioned restriction is a move that is not only convenient, but also recommendable. Therefore, in what follows I will limit myself to employing this model for inquiring into how can be the proponent of philosophical anarchism reconciled with *political* authority.

[75] Raz, *Authority, Law and Morality*, p. 212.

3.4 Beyond the Razian model: some suggestions

Recall from the first part that the major claim advanced by the anarchist is that the authoritative command of a certain course of action cannot *per se* count as a moral reason for action. This claim has a double source. On the one hand, it is fueled by the view that one's submission to government's directives constitutes a breach of one's autonomy. On the other hand, it arises from the belief that even the mere cooperation with a government would lead to results that are morally questionable, or at least morally less significant than acting upon one's own reasons. The second of these sources of concern for the anarchist seems to be readily eliminate by recalling that, according to the normal justification thesis from the Razian model, the anarchist will better further the reasons that apply to herself by following the directives of authority. However, the problem is not so simple with the first of her concerns. As a consequence of the work done by the preemptive thesis, the anarchist must be shown that she has a reason to cooperate with the government by accepting its directives even when by so doing she acts against her best judgment of the case. This aspect becomes further complicated when her best judgment indicates not only that the directives are morally debatable, but that they are plainly morally wrong. As I suggested already in the first part, it is precisely the task of showing the philosophical anarchist why she has reasons to abide by the directives of the authority even in this case that one has to pursue in order to plausibly argue against her position. In this part of the text I try to deal with this task.

As I said, the normal justification thesis seems to offer a good starting point for meeting the anarchist challenge by stating that authoritative directives increase the chances that authority's subjects better comply with the reasons that apply to them. This

statement, however, is in need of an adequate defense, and this can be done by looking at what can be its plausible ground. The more general question we have to answer is 'what is it that makes it the case for an authority **A** that it is a legitimate authority for its subject **S** if and only if the reasons that support complying with **A**'s directives establish them as justifiably preempting the reasons that would otherwise determine **S**'s actions and, by doing so, increases **S**'s chances to better comply with these reasons'?

One of the available answers grounds authority's legitimacy in its having a greater expertise in a certain area. The argument from expertise is usefully discussed by R. B. Friedman in his *On the Concept of Authority in Political Philosophy*. Friedman thinks that an essential feature of the concept of authority is the existence of "the recognition and acceptance of certain criteria for designating who is to possess this kind of influence".[76] This feature of authority is present, Friedman tells us, both in the case of 'an authority' and to the case of 'in authority', the two members of a well-known distinction that he introduces. However, the argument from expertise is characteristic only of the claim at being 'an authority'. This claim rests on two presumptions: a) that there is an inequality of knowledge and insight between the one who possesses authority and the one who enters in relationship with her – a claim of belief, and b) that such a superior knowledge and privileged insight is possible – an epistemological claim. Using Friedman's clarification for my present argument, let me first must notice that the claim to possess expert authority amounts to saying that '**A** is a legitimate expert authority for its subject **S** if and only if the reasons that support complying with **A**'s directives establish them as justifiably preempting the reasons that would otherwise determine **S**'s *beliefs* and, by doing so, increases **S**'s chances to better comply with these reasons'.

[76] R. B. Friedman, *On the Concept of Authority in Political Philosophy*, p. 71.

This, taken together with the announced limitation to the study of political *authority*, can seem to constitute a good enough reason to reject the use of the argument from expertise. On the one hand, the authority of an expert, usually embodied by someone whose assertions (belonging to her area of expertise) are considered true by the members of a given group, is both an essential feature of many areas from the realm of theoretical reasoning and a most useful and welcome presence in a world in which good performance in an increasing number of aspects from most of individuals' everyday life seems to depend on their having access to expert opinions. On the other hand, it seems that expert authority has no role to play into a plausible justification of political authority, for the commonly accepted reason that political authorities that advance a plausible claim to legitimacy can only claim to regulate the behavior of their subjects, while none but the most authoritarian ones claim also to regulate their subject's beliefs.

However, I do not think that we should move too quickly here. Remember that, according to Friedman, the proper characterization of 'an authority' includes both a claim of belief and an epistemic claim. If one places a strong focus on the epistemic claim, looking for an ally in the use of the concept of 'political truth', and at the same time attempts to justify political authority's intervention into the realm of its subjects beliefs by claiming that the authority is able to better lead them toward the realization of the alleged truth, then maybe one can argue for a place that authoritarianism should occupy within the justification of political authority. Consider the general form of the authoritarian argument, as put forward by David Estlund's in his *Making Truth Safe for Democracy*. Estlund starts by observing that democrats will always want to challenge the inference from the unequal distribution of political wisdom to the superiority of

authoritarian political institutions. In his view, the most complex version of the authoritarian argument is what he calls "Normative Epistemic Authoritarianism". As such, this argument implies the endorsement of the following three tenets:

1) *The Cognitivist Tenet*: Normative political claims (at least often) are true or false.

2) *The Elitist Tenet*: Some (relatively few) people know the normative political truth significantly better than others.

3) *The Authoritarian Tenet*: The normative political knowledge of those who know is a strong moral reason for their holding political power.

Among the three ways of challenging the mentioned inference that Estlund identifies, the most interesting one consists in denying that there is such a thing as a need for normative political truth (i.e. it denies the cognitivist tenet – the first premise in the argument for authoritarianism). Although Estlund seems to imply that there are no sufficient reasons for an *a priori* rejection of the claim that the concept of 'political truth' is a meaningful one, he claims that the authoritarian argument can be resisted by looking at an important epistemic difficulty inherent in this argument. In his own words, this difficulty can be expressed by the question "who will know the knowers?" The reason why the proponent of the authoritarian argument is not able to properly answer this question is that, in fact, this question reveals a fourth tenet that the argument for authoritarianism implies, namely:

4) *The second-order epistemic tenet*: The knowers can be known by sufficiently many nonknowers to empower them, and to practically and morally legitimate their power.

As such, the fourth tenet seems to annihilate people's need for political expertise (i.e. the only 'need' that might confer legitimacy to one's claim for a privileged access to political authority). In Estlund's formulation, this boils down to the observation that "even if some have knowledge, others have no way of knowing this unless they can know the same thing by independent means, in which case they have no use for political expertise".[77]

I find Estlund's ingenious argument convincing enough. Still, we must be clear about what exactly it convinces us of. The argument only shows that, since one does not have a special need for political expertise, alleged 'political experts' cannot advance a special claim for holding political power. It does not help us decide whether, as a matter of principle, political authority should or should not aim at being justified by or associated with political expertise. The following state of affairs seems both possible and justified according to Estlund's argument. Suppose that, once in power and representing a state's political authority, a group of 'political experts' often convince (some of) the subjects of that authority that their initial positions on certain matters political are mistaken, to the effect that the subjects accept the experts' conclusions even when they conflict with their initial reasons. Although a certain authoritative relationship will obtain in this case, it will not have the characteristics of the Razian authority anymore. If the subjects are indeed convinced by the authority's directives, then they are in fact acting upon reasons that they regard as valid. Therefore, the directives will not replace subjects'

[77] Estlund, *Making Truth Safe for Democracy*, p. 84.

judgments on the balance of reasons. Hence, Raz's preemptive thesis will not obtain. In this case, it seems that a political authority's resorting to 'political experts' can lead to dissolving the authority relation itself.

The conclusion we are entitled to draw is that a political authority conceived in accordance with the argument from expertise can only retain its authoritative character by giving up the aim of being fully justified to its subjects. In order to avoid this conclusion, it seems that a necessary step that we must take is to renounce the idea that the normal justification of political authority can properly include the belief in the authority's superior ability to practical reasoning. This move involves the search for a different version of the normal justification thesis.

Before going on, some preliminary remarks are in order. To begin with, it is important to notice that the promotion of social values always require a collective effort. Promotion of such values provides both the representatives of the political authority and its subjects with reasons for action. To better understand why this is so, we can refer to such social values as being those kinds of goods in the constitution of which a duty is incorporated. In his *Liberating duties*, Raz argues that, on the one hand, "duties may be internally related to their justifying goods", and, on the other hand, that there are "some activities and relationships which cannot be specified except by reference to duties", some of these latter ones being "intrinsically good".[78] It is something like this that I have in mind in claiming that the promotion of social values offers all those who participate in political societies reasons for cooperation. Moreover, it is also plausible that the deontological principles present in a given society are taken into consideration by all members of that society, even if directly in the case of those who uphold them and

[78] Raz, *Liberating Duties*, p. 41.

merely indirectly in the case of would-be public officials who should strive not to obstruct them. Consequently, it follows that both types of elements that are to be found at the basis of individuals' conceptions of the public good provide the individuals in cause with moral reasons for action. Also, since in this way authority's directives will be based on a conception of the public good which, in turn, provides reasons for action for its subjects, it can be said that the authority grounds its directives in reasons that already independently apply to its subjects.

Up to this point, I tried to look at how can the dependence thesis from Raz's model of authority be reinterpreted and reinforced by employing the above-mentioned assumption. Although this may not come as a surprise, an important consequence arises, namely that the members of a political community can be expected to cooperate in order to realize certain shared social values. However, since in highly heterogeneous societies one can expect to encounter differing, or even contradicting conceptions of the public good, the only use we can find for the mentioned consequence is that it renders plausible the circumstance that individuals would prefer to cooperate for the implementation of a unique conception of the public good to the social disorder in which everyone's strive to realize her own conception of the public good would result. Still, we are left with the task of looking at whether the preemptive and the normal justification theses are also satisfied. In what the preemptive thesis is concerned, the answer seems to be affirmative since the directives of the considered political authority will have to preempt some of the reasons held by some of the authority's subjects, namely those ones that will conflict with the conception of the public good that will be finally adopted.

In what the normal justification thesis is concerned, the things are not clear-cut. While it is true that some subjects are likely to better advance their conception of the public good by acting upon the provided reason to cooperate, it is also true that there will be some other subjects whose advance of *their* conception of the public good will be so far from optimal, that they can be said to acquire a new reason for action, namely one to free-ride on the cooperative efforts of their fellow citizens. It then appears that, if we consider things from the point of view of individual rationality, the normal justification thesis fails to obtain. A tempting reply is to contend that the entire point of practical reasoning is to enable us to get more of what we want or to avoid more of what we want to avoid, in order to be able to conclude that cooperating must be what practical reason requires.[79] However, by so doing, we would implicitly assert that the reason to cooperate must be seen as always being stronger than any other reasons for action individuals might have, and this is in no way granted by my present argument.[80]

However, even if it seems to follow that the members of a political community can be plausibly expected to cooperate in order to realize certain shared social values, the argument above runs into the following problem. Since in highly heterogeneous societies one can expect to encounter differing, diverging, or even contradicting conceptions of the public good, the only use we can find for this claim is that it renders plausible the circumstance that individuals would prefer the alternative of cooperating for the implementation of a unique conception of the public good to the social disorder in which everyone's strive to realize her own conception of the public good would result. Furthermore, while it is true that some individuals are likely to better advance their

[79] This is Kurt Baier's position. See Kurt Baier, *The Moral Point of View*, p. 301.
[80] I briefly touch on this issue in the next section.

conception by acting upon the provided reason to cooperate, it is also true that there will be some other subjects whose advance of *their* conception of the public good will be so far from optimal, that they can be said to acquire a new reason for action, namely one to free-ride on the cooperative efforts of their fellow citizens. It then appears that, if we consider things strictly from the point of view of individual rationality, any reason that individuals can possibly have for cooperating is in a constant danger of being overridden over time. In order to deal with this point, in the next chapter I will focus on a distinction proposed by Christopher McMahon between what he calls "the principle of individual rationality" and "the principle of collective rationality". According to him, even if individuals do not have a reason to cooperate as long as they consider the matter from the standpoint of individual rationality, they might nevertheless have a reason to cooperate if they consider the matter from the standpoint of collective rationality.

Chapter 4
Collective rationality

4.1 The principle of individual rationality and the principle of collective rationality

In *Authority and Democracy*, McMahon proposes a distinction between what he calls "the principle of individual rationality" and "the principle of collective rationality". Before moving on, let me quote his formulation of the two principles.

> **The Principle of Individual Rationality (PIR):** One has reason to contribute to a cooperative venture that produces something that one regards as good if the incremental increase in the value of this good that will be created by one's contribution exceeds the cost to one of contributing.

> **The Principle of Collective Rationality (PCR):** One has reason to contribute to a cooperative venture that produces something that one regards as good if its total value to one when one's contribution is added to those of the others who have contributed or will contribute exceeds the cost to one of contributing.[81]

Let me focus on the character and the function that McMahon ascribes to the second of them since, as it will shortly become obvious, it will play an important role. Two of the characteristics of the principle of collective rationality (hereafter PCR) are outstanding. The first one is that the PCR is not a moral principle. This should be

[81] McMahon, *Authority and Democracy*, pp. 103–105.

understood in the sense that this principle singles out whatever individuals have good reasons to do, irrespective of their moral beliefs concerning a given state of affairs. The second aspect is that the rationality alluded to by the PCR is not "the rationality of a group understood as a distinct entity".[82] Rather, it merely represents "a different way for individuals to process the rational import of the considerations that they acknowledge as reasons for action".[83] In what its actual functioning is concerned, the PCR will only provide a reason to cooperate in those cases when enough others will cooperate to the production of a good whose value exceeds for one the cost of one's cooperation. This further implies that whenever the cooperative enterprise meeting the condition specified by the PCR does not yet exist, the principle will only provide a reason to cooperate if the '*assurance problem*' can be solved, i.e. only if there are reasons for each participant to believe that enough others will cooperate.

Returning now to the main argument, it becomes obvious that the question we have to answer in order to see if we can or cannot justify political authority to the proponent of philosophical anarchism is whether rational moral agents will act upon the PIR or, alternatively, upon the PCR. On the one hand, as McMahon notices, when each is

[82] McMahon, *Authority and Democracy*, p. 104.

[83] McMahon, *Authority and Democracy*, p. 104. It is worth stressing that, while McMahon offers an account of collective rationality, he rejects the idea of collective agency. For more on this point, see McMahon, *Shared Agency and Rational Cooperation*. Even though McMahon acknowledges the importance of the phenomenon of shared action in our lives, he disagrees with much of the recent work that focuses on this phenomenon, insisting that the type of agency displayed in shared action is to be understood by reference to individuals and only derivatively by reference to groups. More specifically, he argues that shared agency can be understood as rational cooperation among individuals and that this account of shared agency is able to explain the characteristic features of shared action. Unlike other authors who attempt to provide an account of shared agency by focusing initially on small groups and arguing then that the account offered can be extended to large groups (such as states), McMahon begins by trying to prove that, in the case of large groups, his theory of rational cooperation can successfully accommodate the characteristic features of shared agency, and then moves on to argue that the account he provides can also be applied to smaller, informal groups. Arguably, one of the advantages of the latter strategy is that it renders the attempt to account for legitimate political authority less problematic. For a detailed discussion of some of the difficulties involved in applying Gilbert's account of joint commitments to the political sphere, see *Appendix*.

seen as an individual rational agent and "when the possibility of being a free-rider is admitted, each will correctly judge that the reasons which she regards as applicable will be better served by rejecting the government's authority".[84] On the other hand, since each finds herself in a situation that has the character of a multi-person prisoner's dilemma, the reverse might be the case, i.e. individuals would arguably consider that collective rationality takes precedence over individual rationality. The latter might be true primarily due to the already discussed circumstance that cooperation with authority prevents the conflict of diverging conceptions of the public good result into social chaos.

McMahon's suggestion is that "representing cooperation with a government as a way of solving prisoner's dilemma problems may be tantamount to a refutation of anarchism".[85] This conclusion is only a momentary one, and there are various important difficulties to be considered. I will discuss some of these difficulties later on. For now, I will limit the discussion to two points. First, I will briefly stress some of the advantages of using this model of justifying political authority for dealing with the anarchist challenge, as compared to other ways of objecting the anarchist position. Secondly, I will consider the relationship between this model and the problem of political obligation.

4.2 Advantages of employing the principle of collective rationality

Let me now turn to the question of what are the main advantages of employing an account based on the PCR. A first advantage consists in the fact that this account does not

[84] McMahon, *Autonomy and Authority*, p. 314.
[85] McMahon, *Autonomy and Authority*, p. 314.

refer to the problematic existence of an independent moral reason to obey the law. Since the argument only seems to appeal to reasons for action provided by the anarchist's own conception of the public good, the claim that these reasons are just as problematic as the moral reason to obey the law seems no longer available to the anarchist. Another important advantage of this model is that it offers a plausible construal of the normal justification thesis, one that dispenses with the problematic requirement advanced by the argument from expertise. On the proposed view, what matters from the point of view of the subjects is that authority offers a solution to the prisoner's dilemma situation that emerges from the confrontation between different conceptions of the public good.

A further advantage seems of to be that of offering a more plausible construal of the relationship between a political authority and its subjects. On the proposed view, all that matters from the point of view of the subjects is that the authority to which they are subjected offers a solution to the prisoner's dilemma situation that emerges from the confrontation between different conceptions of the public good. Thus, individuals might both hold that they have reasons to cooperate with political authority and acknowledge that political authority to which they subject themselves might be mistaken in its consideration of different aspects of the public good.

Once we suppose that the attempted reply to the anarchist challenge is plausible, the question is in what sense and to what extent does this argument entail the existence of political obligations. Notice that if the members of a political society viewed as a cooperative enterprise find living within a political society preferable to the non-society alternative, then it can be safely concluded that they have sufficient reasons to cooperate towards the preservation of the societal arrangement.

Recall that in his *Moral Principles and Political Obligations*, A. John Simmons advanced a claim that is now regarded as a definitive feature of the problem of political obligation, namely 'the particularity requirement'. In what political obligation is concerned, Simmons argues, "we are only interested in those moral requirements which bound an individual to one particular political community, set of political institutions etc."[86] Employing the PCR places a somehow different emphasis on this requirement, while granting its skeptical implications. In other words, although it admits that not every member of a political society will have enough reasons to obey all and only its laws, the PCR arguably requires a greater level of compliance with the law than do the moral principles discussed and rejected by Simmons, viewed as underlined by the PIR.

4.3 Substantive moral principles and the reason to contribute to cooperative schemes

4.3.1 The problem of morally motivated defection

Given that McMahon rests his case for the idea that a formal principle is preferable to a substantive moral principle as a characterization of the requirement to contribute to mutually beneficial cooperative schemes on the claim that the PCR does a better job in characterizing this requirement than the principle of fairness, we need to take a closer look at McMahon's discussion of the principle of fairness.

The principle of fairness holds that, insofar as one has voluntarily accepted the benefits of a cooperative scheme, one is under a prima facie obligation to contribute to

[86] Simmons, *Moral Principles and Political Obligations*, p. 31.

the provision of these benefits, provided that the cooperative scheme in question is just or fair. The fairness-based approach to cooperation has enjoyed a fair amount of popularity over the last decades.[87] Yet, this approach has also been the subject of much criticism, due mainly to the stipulation that benefits must be voluntarily accepted. I will come back to the so-called "voluntariness condition" later on. At this point, let me just stress that McMahon's argument to the effect that the principle of fairness fails to underwrite the requirement to contribute to mutually beneficial cooperative schemes does not depend on whether benefits are voluntarily accepted.

According to McMahon, there are two main reasons why the requirement to contribute to cooperative schemes, when one would be better off[88] by acting as a free

[87] For developing the principle of fairness, see H. L. A. Hart, *Are There Any Natural Rights*? (1955). See also John Rawls, *Legal Obligation and the Duty of Fair Play* (1964) and *A Theory of Justice* (1971), pp. 108–114. (McMahon uses Rawls's formulation of the principle, which makes the fairness of a cooperative scheme a condition of there being an obligation to contribute to it. That the principle of fairness includes this proviso is crucial to McMahon's argument to the effect that the PCR can easily accommodate our basic intuitions about fairness. See note below.) For defending a fairness account of political obligation, see Richardson Arneson, *The Principle of Fairness and Free-Rider Problems* (1982), George Klosko, *The Principle of Fairness and Political Obligation* (1992) and *Fixed Content of Political Obligations* (1998), and Richard Dagger, *Membership, Fair Play, and Political Obligation* (2000). For criticizing the fairness account of political obligation, see Robert Nozick, *Anarchy State and Utopia* (1974), pp. 90–95, John Simmons, *Moral Principles and Political Obligations* (1979), pp. 101–142, *The Principle of Fair Play* (2001*a*), and *Fair Play and Political Obligation: Twenty Years Later* (2001*b*), and Daniel McDermott, *Fair-Play Obligations* (2004).

[88] A qualification may be needed here. As it will soon become manifest, McMahon claims that the requirement to contribute to cooperative schemes should not be understood as a requirement to promote the value of fairness precisely because one can be *morally* better off by acting as a free rider (i.e. one can sometimes better promote one's moral values by acting as a free rider). However, to hold that, in trying to determine whether cooperative behaviour should be seen as grounded in the PCR or in the principle of fairness, one ought to look for the best characterization of the requirement to contribute to cooperative schemes, when one would be *morally* better off by acting as a free rider, would mean to unduly simplify matters. This is because the aforementioned formulation excludes the possibility that one could choose to receive the benefits of cooperation without contributing oneself even when no moral value is promoted by doing so. Yet, this is precisely the type of cases our basic intuitions about fairness tell against. At this juncture, a critic of McMahon's theory can argue that, even if we are inclined to agree with McMahon that the dispositions of cooperatively disposed people are best captured by the PCR as long as we focus on cases when one could better promote one's moral values by acting as a free rider, this will certainly not be the case if we focus on cases when no moral value would be promoted by free riding. Arguably, it is the principle of fairness that provides a reason to refrain from free riding in the latter type of cases. McMahon would most likely answer this objection along the following lines. One of the advantages of the PCR is that it can accommodate our basic intuitions about fairness. (See, for instance, McMahon, *Collective Rationality and Collective Reasoning*, pp. 26–27.) According to McMahon, an individual has reason to make her

rider, is not to be viewed as a requirement to promote the value of fairness. First, when individuals have *moral* reasons for declining to contribute to a cooperative scheme, they may judge that the reason of fairness for contributing is overridden by the moral reason for defecting. However, McMahon points out that, if many individuals choose to defect, this could lead to a state of affairs that each of them judges to be worse, in light of her own moral values, than the state of affairs that would result if all of them made their assigned contributions. Second, McMahon contends that the fairness-based approach to cooperation is rendered problematic by the assumption that cooperators have moral concerns. According to him, moral concerns cannot be viewed as normatively akin to interests, and therefore, the notion of fairness cannot appropriately be said to govern the realization of these concerns. In what follows, I will consider each of these worries in turn.

Let me start by focusing on McMahon's claim that the substantive moral reason for contributing supplied by the principle of fairness can be overridden by a moral reason for defecting. McMahon emphasizes that this worry is especially prominent in political contexts, when the principle of fairness is invoked to establish that citizens have a

assigned contribution to a cooperative scheme under the PCR if the cooperative outcome would be preferable, in light of her own values, to the noncooperative outcome. McMahon argues that individuals who care about morality may find that the cooperative outcome constitutes an improvement over the noncooperative outcome for different reasons, including reasons of fairness. Thus, an individual who cares strongly about fairness may judge, under the PCR, that contributing to a cooperative scheme is required because the outcome produced by scheme would be *fairer* than the noncooperative outcome. In other words, the PCR can accommodate the intuition that an individual can be required to contribute to a cooperative scheme only if the scheme in question is a fair one. Yet, to claim that a cooperative scheme is fair is to claim that the distribution of burdens and benefits within the scheme is fair. Thus, we can say that an individual who cares strongly about fairness may judge, under the PCR, that she is required to contribute to a cooperative scheme because this is what a fair distribution of burdens and benefits requires. In short, McMahon can maintain that the PCR has no trouble accounting for cases about which we may, at first glance, think that only the principle of fairness can account for. If McMahon is right in claiming also that the PCR enables us to explain why individuals refrain from free riding in cases when the principle of fairness fails to offer an adequate explanation, then it seems that the PCR can suitably replace the principle of fairness as a characterization of the requirement to contribute to mutually beneficial cooperative schemes.

political obligation to contribute to the maintenance of the state.[89] In the political case, contribution takes the form of obeying all the laws of the state. In order to justify political obligation, one can invoke other substantive moral principles (such as the principle of consent or the natural duty of justice). It should be kept in mind that McMahon's argument to the effect that the principle of fairness fails to underwrite cooperative behavior applies to any principle that aims to provide a substantive moral reason for contributing to a cooperative scheme. McMahon points out that such principles work within the framework of individual rationality. According to the PIR, in order to decide whether to contribute to a certain cooperative scheme, one has to compare "the difference that one can make to the value of the world (as one understands this) by [contributing] with the difference that one can make by using in some other way the resources at issue".[90] Substantive moral principles assign negative moral value to declining to contribute, e.g. free riding is viewed as *unfair*. Thus, if an individual who has to decide whether to make her assigned contribution is tempted not to do so on the basis of a calculus of self-interest, considerations of fairness are supposed to tip the balance in favor of contributing.

Before going further, it is important to emphasize that while McMahon's account of rationality shares with utility theory the idea that an agent has best reason to perform that action from the set of available actions which is correlated with the outcome that has the greatest value (from the agent's standpoint), it also diverges from utility theory in a significant respect.[91] According to utility theory, choice is determined by preference. By contrast, McMahon's account of rationality gives pride of place to principles of value. In

[89] McMahon, *Reply to Gaus, Richardson and Weber*, p. 201.
[90] McMahon, *Collective Rationality and Collective Reasoning*, p. 16.
[91] McMahon, *Collective Rationality and Collective Reasoning*, pp. 6–7.

other words, according to McMahon the desirability of the available outcomes is determined by the agent's principles of value, and not by her preferences. This point will prove to be important at various stages in the discussion to follow.[92]

Going back to the issue of whether cooperative behavior should be seen as grounded in a substantive moral principle, it should be noted that McMahon grants that the principle of fairness and the PIR direct contribution to a cooperative scheme when the agent's reasons for defecting are nonmoral. An agent who accepts the principle of fairness will usually judge that considerations of fairness that support contribution override self-regarding considerations that support noncontribution. However, McMahon contends that the situation is different when the agent's reasons for defecting from cooperation are moral. In such cases, the reason of fairness for contributing must compete with the moral reason for defecting, and there is no guarantee that the former will override the latter. Admittedly, an agent might judge that the moral good she can realize by using in a different way the resources that constitute her share of the costs of the cooperative enterprise outweighs both the incremental loss in the moral gains that she regards as the result of the cooperative endeavor and the reason of fairness for contributing.

In order to establish that the principle of fairness can reliably block morally motivated defection, one would have to show that considerations of fairness are strong enough to outweigh in all circumstances moral considerations for defecting. However, McMahon draws attention to the fact that, when it comes to large scale cooperative

[92] McMahon's account of rationality further diverges from utility theory in rejecting the assumption that the ranking of possible outcomes is complete and transitive. McMahon's account can therefore make room for evaluative incommensurabilities. See McMahon, *Collective Rationality and Collective Reasoning*, pp. 7, 30–32.

schemes, the unfairness brought into the world by an individual's defecting from the scheme is not, on a scale of unfairness, very great. Yet, an individual might achieve a great deal (in terms of her moral values) by defecting. An example will serve to drive this point home.[93] Suppose that an individual who owns a car has to decide whether to bear her fair share of the costs of a pollution-control scheme. If the individual in question is one of several million car owners in a given urban area, then the unfairness displayed by her acting as a free rider on the contributions of other car owners is not very great. She might, however, use the resources that constitute her share of the costs of the pollution-control endeavor to accomplish something of significant value (from her point of view), e.g. to save money in order to fund higher education for her children. To show that not all conflicts between a moral reason for contributing and a moral reason for defecting will be resolved in favor of the former, McMahon asks us to compare the unfairness displayed by an individual who free rides on the cooperative efforts of other car owners with the unfairness displayed by a parent who funds higher education for her son but not for her daughter.

In cases like the one discussed above, the PIR in conjunction with considerations of fairness do not direct contribution. According to the PIR, it is rational in such cases to act as a free rider. Nevertheless, if many individuals act as the PIR and the principle of fairness dictate, they will end up with a Pareto suboptimal outcome (i.e. a situation that each of them judges to be worse, in light of her own moral values, than the situation that would result if all of them made their assigned contributions). McMahon argues that the PCR avoids this problem. As he stresses:

[93] McMahon, *Collective Rationality and Collective Reasoning*, p. 17.

> When an agent faces a choice whether to contribute to a cooperative scheme, the PIR directs him to compare the value of the outcomes produced by two different actions, contributing and declining to contribute, considered as individual events. The PCR, by contrast, directs him to compare the value of outcomes produced by two different *combinations* of actions, of which contributing and declining to contribute constitute parts.[94]

The PCR does not threaten the achievement of Pareto optimality because, unlike the PIR, it directs one to compare the value to one of the outcome of the scheme, when one's contribution is added to the contributions made by others, with the value to one of the outcome in which everyone defects from scheme, and thus it directs contribution.[95]

4.3.2 Fairness and the demotion of moral concerns to interests

There is, however, a deeper problem with claiming that the requirement to contribute to mutually beneficial cooperative schemes should be understood as grounded in the principle of fairness.[96] This problem arises when the cooperators are assumed to have moral concerns. McMahon points out that the notion of fairness typically applies to the distribution of benefits and burdens among the members of a group. Yet, he argues that it is far from clear whether it is appropriate to conceive of the promotion of individuals' different moral values as matter of distributing benefits to them. Let me detail.

[94] McMahon, *Collective Rationality and Collective Reasoning*, p. 18 (emphasis in original).
[95] McMahon, *Collective Rationality and Collective Reasoning*, pp. 21–22.
[96] McMahon, *Collective Rationality and Collective Reasoning*, p. 20.

McMahon points out that in a pluralistic society it is inevitable that the moral concerns of some individuals are more fully realized, while the moral concerns of other individuals are less fully realized. It is, however, doubtful that this is unfair. McMahon rightly emphasizes that "those whose concerns are frustrated will regard the world as getting morally worse (by their lights), and they will be justified in doing what they can to prevent this, but not, it seems because the deterioration is unfair to them personally"[97]. Now, granting that McMahon is right in claiming that employing the notion of fairness is inappropriate when we consider the extent to which the moral concerns of different individuals are realized, what follows is that the complaint against those who defect for moral reasons from a cooperative enterprise that produces moral gains is not one of unfairness. Still, there is a sense in which those who defect act contrary to reason. This is because, if many chose to defect, the moral gains of cooperation would be lost, and each would regard the resulting state of affairs as morally worse than the state of affairs that would have resulted if everyone had contributed. The upshot of all these considerations is that, in order to account for cooperative behavior, we need to look further than the principle of fairness. According to McMahon, it is the PCR that provides a reason to contribute.

McMahon's suggestion is that, once we assume that cooperators have moral concerns, the issue of whether cooperative behavior on the part of individuals with such concerns can be viewed as grounded in the principle of fairness boils down to the issue of whether it is appropriate to regard moral concerns as normatively akin to interests. This is because fairness is typically invoked to resolve conflicts of interests.[98] Thus, whether one

[97] McMahon, *Collective Rationality and Collective Reasoning*, p. 20.
[98] McMahon, *Collective Rationality and Collective Reasoning*, pp. 20–21, 40–42, 94–95.

agrees with the claim that the notion of fairness cannot appropriately be applied in contexts when what is at stake is the realization of individuals' different moral concerns depends on whether one considers the demotion of moral concerns to the status of mere interests to be problematic.[99]

It is worth noting that, in a more recent writing, McMahon offers a detailed discussion of the different aspects of political cooperation that are governed by the notion of fairness.[100] In McMahon's view, fairness is to be understood "as a matter of appropriate concession".[101] He makes a distinction between narrow and broad fairness. As he puts it, narrow fairness is concerned with mediating between the claims of desert or need that can be made on behalf of the members of a group. Thus, narrow fairness is concerned with the distribution of benefits and burdens in a group. Yet, McMahon rightly points out that, when the members of a modern political society address the question of how to organize political cooperation, they also bring to bear considerations that concern only indirectly the distribution of benefits and burdens among individuals. These considerations (e.g. the preservation of social peace, the upholding of the rule of law, the promotion of social prosperity, the fostering of the value of community, the advancement of knowledge etc.) are labeled "morally important social values". According to McMahon, whenever the promotion of morally important social values is held to require a particular pattern of concessions, the notion of fairness is used in the broad sense. McMahon stresses that the requirements of narrow and broad fairness can conflict. As an

[99] For holding that a social order in which the value of fairness plays a central role cannot be too bad, even if it requires the demotion of values other than fairness to the status of interests, see Michael Weber, *The Reason to Contribute to Cooperative Schemes: An Examination of Christopher McMahon's "Collective Rationality and Collective Reasoning"*, p. 178.
[100] McMahon, *Reasonable Disagreement: A Theory of Political Morality* (2009), pp. 71–92.
[101] McMahon, *Reasonable Disagreement: A Theory of Political Morality*, p. 85.

illustration of this point, consider the case of affirmative action.[102] Admittedly, narrow fairness demands that all qualified candidates, regardless of gender, race or ethnicity, are given equal employment opportunity. However, since the preferential hiring of members of previously excluded groups would arguably foster the value of community, one can claim that broad fairness demands that we adopt this policy.

Note, however, that the distinction between narrow and broad fairness does not help to settle the issue of whether applying the notion of fairness is appropriate in contexts when what is at stake is the realization of individuals' different moral concerns. Yet, given that fairness is largely understood as matter of appropriate concession, a question that legitimately arises is to what extent considerations of fairness are relevant to determining under what circumstances it is appropriate to make concessions from one's preferred moral view. I will come back to this question later on. For now, let me just stress that the unwavering disposition to make and to seek concessions, which is manifested, according to McMahon, by cooperatively disposed people, is expressed not only in connection with issues that are appropriately governed by fairness, narrowly construed, but also in connection with issues that fall within the scope of fairness, broadly construed.

4.4 McMahon and Gaus on fairness

As we have seen, McMahon argues that fairness cannot legitimately be held to govern the realization of individuals' divergent moral concerns. This is because, in his

[102] McMahon, *Reasonable Disagreement: A Theory of Political Morality*, p. 87.

view, there is something deeply problematic about the suggestion that moral concerns have the same normative status as interests, and therefore, moral conflicts ought to be resolved the same way that conflicts of interests are resolved. It is worth mentioning that, in a different context, a similar worry is raised by Gerald Gaus. In his *The Order of Public Reason*, Gaus revisits the idea that compromise is at the heart of public justification. As he points out, that reasonable persons who want to live together in a mutually beneficial social order have to exercise the virtue of meeting the other halfway seems to be a fact that is hardly worth disputing. To stick to Gaus's example, if they have to decide how to divide the product of joint labor in the making of which it is difficult to track individual contributions, reasonable persons will show a disposition to concede to others their share, manifesting thus a disposition to compromise. In fact, one can say that showing unwillingness to compromise because of one's superior bargaining power is part of what it means to be "unreasonable". However, Gaus argues that if we think of public justification as being concerned with whether a given rule can be endorsed from one's evaluative viewpoint, to maintain that public justification is fundamentally about compromise is suspect. As he puts it:

> To say that public justification involves splitting the difference between what a religious person believes is justified and what an ardent secularist holds to be supposes that living according to one's evaluative standards is like claiming a share of a common product, to be negotiated away.[103]

Taking his cue from Rawls's *Political Liberalism*, Gaus emphasizes that deliberation among reasonable persons is informed by a concern to accommodate the fact of

[103] Gerald Gaus, *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, pp. 331–332.

pluralism. He points out that a reasonable person knows that others have different evaluative standards and understands that living together under common rules involves accepting the possibility that these rules will not necessarily be those that best satisfy one's evaluative standards. However, he goes on to stress that:

> […] while such accommodation to the reality of pluralism is necessary, this does not show that public justification is centrally about the correct compromise concerning how much the moral rules we live under reflect our basic normative convictions, as if they were pots of money to be divided up, or negotiation aims to be haggled for.[104]

A brief point of clarification is called for. What interests me in this section is whether we can agree with McMahon's claim that, despite common assumptions to the contrary, the principle of fairness does not offer the best characterization of the reason to contribute to mutually beneficial cooperative schemes. It is important to note that, in addressing the issue of how cooperation can be guided by reason, McMahon distinguishes between two aspects of cooperation.[105] The first has to do with whether individuals have sufficient reason to contribute to cooperative schemes, while the second has to do with whether the choice of a given scheme is justified. As McMahon points out, the principle of fairness is commonly thought to play a decisive role in settling both these matters. However, he stresses that in claiming that the PCR can suitably replace the principle of fairness, it is the first aspect of rational cooperation that he has in mind.[106] We have seen that, according to McMahon, there are two main difficulties with holding that the reason to contribute to cooperative schemes is best accounted for by the principle

[104] Gaus, *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, p. 332.
[105] McMahon, *Collective Rationality and Collective Reasoning*, p. 1.
[106] McMahon, *Reply to Gaus, Richardson and Weber*, p. 202.

of fairness. First, a substantive moral principle, such as the principle of fairness, is arguably unable to block morally motivated defection. Second, to insist that individuals who have moral reasons for defecting from a cooperative enterprise must always judge that the reason of fairness for contributing prevails over the reasons for defecting provided by their moral values would mean to insist that individuals should be prepared to demote their moral values to interests. McMahon argues that the PCR avoids these difficulties. It should be noted, however, that McMahon acknowledges that considerations of fairness play an important role in our thinking about cooperation. More specifically, he holds that whether the choice of a particular cooperative scheme can be regarded as justified will depend on whether the scheme in question is a fair one.[107]

A critic of McMahon's approach to rational cooperation can draw attention to the fact that the problem which arises when fairness is invoked in connection with the first aspect of cooperation, namely that individual have to accept the demotion of their moral concerns to interests, surfaces again when fairness is invoked in connection with the second aspect of cooperation. It should be emphasized, however, that McMahon acknowledges that when the choice of a cooperative scheme to be implemented is at issue, the demotion of certain moral concerns to interests may be inevitable.[108] As already mentioned, McMahon's interest is in issues related to rational cooperation among individuals holding divergent moral values. And such individuals might certainly disagree about which scheme to implement based on their moral values. Fairness might

[107] McMahon acknowledges that considerations of fairness play an important role not only in connection to the second aspect of rational cooperation, but also in connection to the first one. McMahon does not deny that individuals who take themselves to have reason to contribute to a cooperative scheme commonly judge that doing otherwise would be unfair to others. He argues nonetheless that our basic intuitions about fairness can easily be accommodated by the PCR. See *supra* note 2.
[108] McMahon, *Collective Rationality and Collective Reasoning*, pp. 40–42, 93–95.

be thought to provide a basis on which such disagreements can be resolved. Presumably, all members of a cooperative group take considerations of fairness to be relevant to the choice of a cooperative scheme. Thus, the choice of a particular scheme can be regarded as appropriate based on considerations of substantive or procedural fairness (i.e. it is because a scheme mediates fairly between competing claims, or because it has been chosen by a fair procedure, that its adoption is deemed appropriate). Yet, even if we leave aside complications having to do with whether the members of a cooperative group do in fact share a common conception of substantive or procedural fairness, invoking fairness as a basis of agreement when cooperators hold divergent moral values is still problematic. McMahon argues that some members of the group will be able to regard the choice of a given scheme as justified only insofar as they judge that fairness outweighs the moral values that support the adoption of a different scheme. In other words, they would have to be prepared to demote the concerns grounded in these values to mere interests. In a nutshell, the problem that McMahon brings to our attention is that, given the assumption that cooperators have divergent moral concerns, attaining a state in which the choice of a scheme is viewed by all as justified seems impossible unless some accept the demotion of their moral concerns to interests. However, it is important to be clear about McMahon's position on this matter. He holds that the demotion of moral concerns to interests is always something regrettable and insists that "it would be desirable if, at each juncture where the expedient presents itself, we could avoid it".[109] What follows is that, in characterizing the reason to contribute in cooperative schemes, the PCR is to be preferred to the principle of fairness because it allows us to avoid demotion.

[109] McMahon, *Collective Rationality and Collective Reasoning*, p. 21.

Going back to Gaus's remarks on the issue of whether compromise is at the heart of public justification, there are two things to notice. First, in stressing that there is something dubious about the suggestion that a reasonable person would have to manifest a fundamental disposition to compromise when it comes to her moral views, Gaus's primary concern is in whether the existence of rules that each can regard as justified ultimately depends on whether reasonable persons manifest such a disposition to compromise. (Here one can ask whether in raising the above-mentioned worry Gaus is interested in what I have referred to, following McMahon, as the second aspect of rational cooperation rather than the first aspect. I address this question below.) Second, Gaus goes to great lengths to demonstrate that the members of a group can choose a rule which can appropriately be regarded as uniquely justified (i.e. it is the rule that all members of the group have reason to endorse), and that this does not at all involve a disposition to compromise.[110] This is an important point of divergence between McMahon's and Gaus's thinking about rational cooperation. As we have seen, McMahon maintains that, whichever scheme is chosen, some members of the cooperative group will be unable to regard the choice as justified unless they are prepared to view fairness as prevailing over their other moral values. (Note, however, that McMahon contends that even if an individual cannot take the choice of a particular scheme to be guided by reason, the PCR can still give her a reason to contribute to the cooperative scheme in question. On his view, it is precisely because convergence on the judgment that a particular scheme is justified is unattainable that a principle like the PCR is needed to direct contribution.) However, it should be kept in mind that, despite their other

[110] Gaus, *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, pp. 332, 389–409.

differences, both McMahon and Gaus urge that, whenever what is at stake is the degree to which individuals' conflicting moral concerns are realized, to insist that individuals must manifest a fundamental disposition to compromise is seriously misguided. This is why each of them is at pains to show that the account he provides does not make the rationality of cooperating depend on individuals' manifesting such a disposition.

Before going any further, it is worth pausing to consider in more detail how Gaus's approach to rational cooperation diverges from McMahon's. McMahon argues that, given the assumption that individuals have divergent moral concerns, attaining a state in which the choice of a cooperative scheme is regarded by all as justified seems impossible unless some accept the demotion of their moral concerns to interests. As already noted, Gaus disagrees on this point. In his terminology, all free and equal persons have conclusive reason to conform to a moral rule in equilibrium. Given that such a rule best satisfies the evaluative standards of each, each has a reason to conform to this rule rather than any alternative. What enables Gaus to reach the conclusion that there is a rule that best satisfies the evaluative standards of each is the idea that an agent who exercises her freedom simultaneously with other agents and aims at coordinating with them has to take into consideration what others are doing. Gaus claims that the freedom of such an agent "is not the freedom of an asocial agent who is free simply when she does what she thinks best regardless of what others do, but the freedom of a social moral agent who considers what her evaluative standards deem is the best thing to do *given* what others

justifiably do on the basis of their own standards".[111] On this view, a rule in equilibrium is a rule that best fulfills one's evaluative standards *given what others are doing*.[112]

Given that McMahon and Gaus start from several common assumptions, it is an interesting question why they reach different conclusions as to whether the members of a group can choose a rule that each can regard as justified. Both these authors assume (a) that individuals have different evaluative standards; (b) that in determining whether a rule is justified, or whether one has a sufficient reason to act on a rule, individuals consult their evaluative standards[113]; and (c) that rationality requires taking into consideration what others are doing.[114] Arguably, it is because (c) is glossed differently by McMahon and Gaus that they arrive at different conclusions as to whether there can be a rule that each will deem justified. Although both these authors insist that taking into account how others act is a requirement of practical rationality, they disagree when it comes to the stage at which this requirement is supposed to operate. On McMahon's view, individuals consult their evaluative standards in order to determine whether adopting a particular scheme is justified. Even if it turns out that one cannot regard the adoption of a particular scheme as justified, the PCR can still provide a reason to contribute to the scheme in question. More specifically, the PCR directs contribution by prompting the individual to

[111] Gaus, *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, p. 402 (emphasis in original).
[112] Gaus, *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, p. 401.
[113] We have already seen that McMahon's account of rationality diverges from standard utility theory in that choice is not determined by an agent's preferences, but by her principles of value. Compare, for instance, Gaus, *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, pp. 395–396. Gaus emphasizes that he takes utility to be a measure of the raking of the available options based on the agent's evaluative standards.
[114] Note further that Gaus writes that: "The choice of a moral rule in a Social Realm of Ends is not a collective "we" choice; the group, *as a group*, does not choose its moral rules". See Gaus, *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, p. 401 (emphasis in original). Compare McMahon, *Authority and Democracy*, p. 104.

consider how the realization of her values would be affected by everyone's defecting.[115] On Gaus's view, the requirement to take into account what others are doing enters the picture at an earlier stage of deliberation. Gaus claims that, in considering the question of whether a scheme can be deemed justified from the point of view of one's evaluative standards, a free and equal member of a cooperative group cannot abstract from the question of whether others judge the scheme to be justified based on their own evaluative standards. This claim goes hand in hand with the view that our normative convictions are not static. Gaus insists that:

> […] our evaluative standards themselves are transformed when we witness how a mutually acceptable cooperative scheme of agency and other rights allows all to follow the deepest convictions while treating their fellows as free and equal moral persons. […] Normative theory is not, as it were, constructed from scratch. We are not in a state of nature, nor are we thinking for the first time what our moral life looks like. We have been formed in a moral order, our standards reflect such a moral order [...].[116]

A detailed discussion of the comparative merits of the two above-mentioned views about practical rationality is beyond the scope of the present chapter. However, let me just add that Gaus maintains that the process by which the members of a group arrive at a publicly justified morality can be understood as a social evolutionary process. In his view, a

[115] For more detail on how actions are supposed to be correlated to valued outcomes under the PCR, see for instance McMahon's discussion on whether the PCR ought to be regarded as a formal moral principle given that this principle seems to involve a form of universalization. See McMahon, *Collective Rationality and Collective Reasoning*, p. 28. McMahon argues that the PCR is best viewed as a nonmoral principle. He draws attention to the fact that the PCR makes less use of the universalization device than it might make. For instance, while the PCR prompts the agent to assign a value to defection by asking what would happen if everyone defected, it does not prompt the agent to assign a value to contribution by assuming counterfactually that everyone contributes. Compare McMahon, *Collective Rationality and Collective Reasoning*, p. 8. See also Gaus, *Once More unto the Breach, My Dear Friends, Once More: McMahon's Attempt to Solve the Paradox of the Prisoner's Dilemma*, p. 168, for highlighting some inadvertencies in McMahon's view on how the PCR correlates actions and valued outcomes.

[116] Gaus, *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, p. 409.

publicly justified moral rule is a Lewisian convention.[117] I will have more to say about Lewisian conventions in the next chapter. In this section, I will limit myself to spelling out the implications of adopting one of the two above-mentioned views about practical rationality or the other.

4.5 Two aspects of rational cooperation

At this juncture, it is also worth noticing that distinguishing between two aspects of rational cooperation is one of the peculiarities of McMahon's approach. As McMahon himself stresses, the fact that he does not take the rationality of contributing to a particular cooperative scheme to depend on whether the scheme in question has been rationally chosen is a key difference between his approach and the one developed by David Gauthier in his *Morals by Agreement*.[118] The main motivation for separating (pace Gauthier) the reason to make one's assigned contribution to a cooperative scheme from considerations that have to do with the optimality and fairness of the scheme in question is that, according to McMahon, an equilibrium solution in a coordination problem need be neither optimal nor fair.[119] McMahon insists that it would be wrong to think that, when agents face ordinary coordination problems, rational cooperation is restricted to creating outcomes that are nearly optimal and fair. It should also be added here that, on McMahon's view, to be cooperatively disposed means to be disposed to treat prisoner's

[117] Gaus, *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, p. 410.
[118] McMahon, *Collective Rationality and Collective Reasoning*, pp. 37–38.
[119] McMahon, *Collective Rationality and Collective Reasoning*, pp. 38–39.

dilemma situations as ordinary coordination problems.[120] This allows McMahon to extend the point made above to prisoner's dilemma situations. And if, as he points out, multi-person prisoner's dilemmas are unavoidable in cooperative contexts, this is not a negligible result.[121]

Whether or not we find fault with the line of reasoning described above, we can agree that anyone who shares McMahon's skepticism about the possibility of choosing a scheme that all members of a large scale cooperative group can regard as justified had better not make the reason to contribute to a scheme depend on whether one can regard the choice of the scheme as appropriate, if any hope to prove that there is ever sufficient reason to participate in cooperative enterprises is to be preserved. As long as these two aspects of rational cooperation are kept separate, a principle such as McMahon's PCR can give one a reason to contribute to a scheme when one cannot take the choice of the scheme in question to be guided by reason. However, if one takes the opposite view, i.e. if one is confident about the possibility of choosing a scheme that all can deem justified, insisting on there being two aspects of rational cooperation seems pointless. (Recall that above I have raised the question whether we should think of Gaus's worry about whether accommodating to the fact of reasonable pluralism essentially involves a disposition to compromise as applying to the second rather than the first aspect of rational cooperation. We are now in a position to answer this question. On Gaus's view, the fact that individuals have divergent moral concerns is not an obstacle in choosing a scheme that all

[120] See McMahon, *Collective Rationality and Collective Reasoning*, p. 21. Note that, according to McMahon, what allows an agent to view a prisoner's dilemma situation as having the structure of an ordinary coordination problem is the PCR. Under the PCR agents assign the same payoff to unilateral defection that they assign to the noncooperative outcome.

[121] See McMahon, *Reply to Gaus, Richardson and Weber*, p. 199. A situation in which cooperation is needed to produce a public good and free riding is a possibility is paradigmatic for multi-person prisoner's dilemmas.

can regard as justified, so there is no need to insist on viewing the two above-mentioned aspects of rational cooperation as distinct. That each can regard the choice of a scheme as appropriate is what provides each with a reason to contribute). What follows from these considerations is that, if an approach to cooperation which aims to prove that the members of a group can choose a scheme that all can deem justified is vindicated, the PCR can be dispensed with. On McMahon's view, it is precisely because convergence on the judgment that a particular scheme is justified seems unattainable that a principle like the PCR is needed to direct contribution. Yet, if McMahon is wrong in thinking that such convergence remains unattainable, there seems to be no need for the PCR to justify contribution. Thus, a critic of McMahon's idea that we should view the PIR as supplemented by the PCR might appeal, apart from general considerations about Pareto optimality and what instrumental rationality requires[122], to considerations aimed at proving the tenability of an approach to cooperation according to which pluralism is not an obstacle to choosing a scheme that all can deem justified.

I have suggested that if, for instance, Gaus's account of how a cooperative group can arrive at a uniquely justified rule (or, to use McMahon's terminology, at a scheme

[122] For this line of criticism, see Gaus, *Once More unto the Breach, My Dear Friends, Once More: McMahon's Attempt to Solve the Paradox of the Prisoner's Dilemma*. Gaus argues that McMahon is wrong in thinking that the PCR would be needed to account for the fact that cooperatively disposed individuals do not defect in a prisoner's dilemma. More specifically, he claims that the PIR can be interpreted in such a way as to provide for cooperation in prisoner's dilemma situations. It is also worth mentioning that Gaus stresses that, on McMahon's view, the PIR is problematic from the perspective of rationality because it threatens Pareto optimality. Yet, he rightly points out that, on this line of reasoning, the PCR should also be viewed as deficient from the point of view of rationality given that it often directs individuals to act in suboptimal ways (i.e. it directs cooperation when one would be better off by acting as a free rider). See Gaus, *Once More unto the Breach, My Dear Friends, Once More: McMahon's Attempt to Solve the Paradox of the Prisoner's Dilemma*, pp. 167–168. It can be added here that McMahon maintains that "the inadequacy of the PIR as a basis for cooperative action consists primarily in the fact that general compliance with its dictates often results in a suboptimal outcome." McMahon, *Collective Rationality and Collective Reasoning*, p. 17. Yet, as we have already seen, his main motivation for distinguishing between two aspects of rational cooperation is that, in his view, there is no reason to see rational cooperation as restricted to creating outcomes that are nearly optimal or fair. A proponent of McMahon's theory must show how the apparent tension between these two claims is to be resolved.

that each deems to be justified) is vindicated, then each member of the group has a sufficient reason to act on that rule (to contribute to that scheme), and so the PCR can be dispensed with. Yet, one might argue that Gaus's account is still prone to the objection raised by McMahon against attempts to account for the reason to make one's assigned contribution to a cooperative scheme by invoking substantive moral considerations. According to this objection, when one has moral reasons for defecting from a scheme, in order to judge that the reason of fairness for contributing prevails over the reasons for defecting provided by values other than fairness one has to be prepared to demote one's moral concerns to mere interests. Since Gaus agrees with McMahon that invoking fairness when what is at issue is the realization of individuals' moral concerns is objectionable, if it turned out that his approach to cooperation has this undesirable consequence, he would have to concede that the reason to contribute to cooperative schemes is better accounted for by a formal principle such as the PCR, which would enable us to avoid this consequence.

Two points should be emphasized here. First, one can respond that if Gaus is right in claiming that the members of a cooperative group can choose a rule that is uniquely justified, then the above-mentioned type of conflict is prevented from occurring. Presumably, this is because to hold that there is such a uniquely justified rule is to hold that this is the (only) rule that everyone has reason to endorse and which can provide the basis of social cooperation among free and equal persons.[123] However, whether this rejoinder can save an account like the one advanced by Gaus will depend on how exactly its details are spelled out. Let me explain.

[123] See Gaus, *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, p. 403. Compare Rawls, *Political Liberalism*, p. 98.

It is important to note that McMahon explicitly addresses the worry that if an approach which tries to justify social cooperation in terms of Rawlsian public reason is vindicated, then the PCR can be dispensed with.[124] He argues, nonetheless, that this worry is unfounded. On the Rawlsian approach, the political conception of justice provides a basis for cooperation that all reasonable persons can accept. The values expressed by the political conception of justice are assumed to prevail over the moral values with which they might come into conflict. Yet, McMahon contends that, in using the PIR to determine whether in a particular case one has sufficient reason to comply with a law, one must compare not only the relevant values considered in the abstract, but also the degree to which various courses of action would affect the realization of these values. McMahon points out that, on the one hand, the fact that an individual fails to comply with a law does not have a significant effect on the degree to which fundamental political values are realized by a given state. On the other hand, however, an individual might achieve a great deal in terms of her moral values by breaking the law. As long as one considers the matter solely from the point of view of the PIR, the fundamental structure of the conflict between the reason that supports compliance with the law and the reason that supports the opposite course of action will remain unchanged, there being no guarantee that the former reason will override the latter. The PCR changes this structure by providing one with a *new* reason to comply with the law. One has sufficient reason under the PCR to comply with the law if one judges cooperation to promote the political conception of justice to be preferable to the best one can achieve in the noncooperative outcome. The upshot is that, even if all reasonable persons have reason to endorse the political conception of justice, this does not justify dispensing with the PCR.

[124] See McMahon, *Collective Rationality and Collective Reasoning*, pp. 78–79.

Thus, if one wishes to defend Gaus's approach to cooperation, one has to show how the approach he advances differs from the one advanced by Rawls. More specifically, one has to show why the above-mentioned type of conflict does not arise on Gaus's approach (or why it has a structure which is different from the one described above).

This brings me to the second point I want to emphasize. One might draw attention to the fact that Gaus is sympathetic to the view that evaluative standards themselves can transform, and argue that such a view is hospitable to the claim that individuals will have no moral reasons to behave uncooperatively. Several features of Gaus's approach are relevant in this context. Gaus maintains that converging on a common moral rule is an instance of the phenomenon of "increasing returns", i.e. the larger the number of individuals who come to embrace a certain moral rule, the more reason others have to embrace it as well.[125] Note also that Gaus stresses that "when living under a justified scheme people come to better appreciate its virtues and become increasingly devoted to it".[126] He points out that "once moral persons live under a justified system of rules most come to embrace it and see that it serves their evaluative standards"[127], and then adds that evaluative standards themselves can transform in light of the moral benefits produced by cooperation. The question is whether on this approach individuals can have moral reasons to behave uncooperatively. An advocate of Gaus's approach might grant that individuals hold moral values that *potentially* dictate defection, but insist that the threat of morally

[125] See Gaus, *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, pp. 398–400.
[126] Gaus, *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, p. 408.
[127] Gaus, *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, p. 408.

motivated defection is reliably blocked by the fact that those who live under a justified scheme become increasingly aware of the moral gains secured by the scheme, and hence, increasingly committed to maintaining it.[128]

What is important for my purposes is that Gaus's approach to cooperation offers a clue to a strategy that can be adopted by those who take seriously McMahon's objections to characterizing the reason to contribute to cooperative schemes in terms of a substantive moral principle but who nevertheless consider that employing such a principle is preferable to employing a formal principle like the PCR. We have seen above that, for instance, the widely-accepted Rawlsian approach to cooperation does not seem to have the resources to respond to McMahon's objections. However, this does not mean that we

[128] The question of whether, on Gaus's approach, individuals can have moral reasons to behave uncooperatively deserves much more careful attention than I am able to devote to it here. One might ask, for instance, whether Gaus's approach licenses the claim that, once a justified system of rules is in place, conflicts of reasons of the type discussed by McMahon are rendered impossible, and whether endorsing the aforementioned claim would be necessary in order to establish that McMahon's PCR can be dispensed with. Note that, according to some commentators, conflicts between a moral reason to contribute and a moral reason to defect are not only possible on Gaus's approach, but also unlikely to be constantly resolved in favor of contribution. For instance, Colin Bird (2011) argues that there is a troublesome ambiguity in Gaus's text between claiming that one has sufficient reason to endorse a rule as morally authoritative within a society and claiming that one has sufficient reason to act on that rule. Bird stresses that Gaus's coordination games aim to show that one has sufficient reason to endorse a moral rule when an interactive convergence develops around it (i.e. when the indeterminacy caused by the diversity of evaluative standpoints is overcome and a given rule is selected as uniquely justified). However, it is far from clear that from the fact that one has sufficient reason to endorse a rule it follows that one has sufficient reason to act on that rule in specific cases. As Bird points out: "By endorsing the rule as authoritative, I (as it were) automatically will that I internalize that rule, such that I accept that I always have sufficient reason to act on it and can anticipate that any dissonance between my 'evaluative standpoint' and the demands of the internalized rule will be within tolerable limits. But this assumption strikes me as illicit and potentially question-begging: what there is convergence upon in the interactive games Gerald Gaus describes is the idea that rule X is morally authoritative within a society. There is not yet, or necessarily, convergence upon the idea that members of that society always have sufficient reason to act on X. And surely it exactly at this point that the problem of reconciling the authority of moral rules with diverse 'evaluative standpoints' emerges." Bird rightly emphasizes that, for instance, an individual who is aware that another rule might have been selected from the optimal eligible set and that her evaluative standpoint is better represented by a different rule, might fail to conclude that she has sufficient reason to act on a rule that requires, in a particular case, acting against what is judged to best from her evaluative standpoint. Note, however, that once we grant that evaluative standards themselves can transform in light of the moral benefits that derive from coordinating on a common set of moral rules, the likelihood of judging that the reason to act on a moral rule that the group has coordinated on is overridden by (moral) reasons that support the opposite course of action, and even the likelihood of such conflicts occurring, is greatly diminished.

should conclude that the PCR is needed to account for the reason to contribute to cooperative schemes. While Gaus's approach has much in common with Rawls's approach (e.g. the idea of public reason is central to both), the former possesses certain features lacked by the latter which may prove to be extremely relevant in the present context. According to Gaus, the Humean/Lewisian conventionalist account has a lot to say about the evolution of morality. The resources provided by this account may help take a significant step in the direction of rehabilitating the idea that the reason to contribute to cooperative schemes should be accounted for by a substantive moral principle.

We have seen that, while McMahon and Gaus share the view that practical rationality requires taking into account what others are doing, they disagree when it comes to the stage at which this requirement is supposed to operate. Let me conclude this section with a few words about the far-reaching implications that derive from adopting one of the above-mentioned views about practical rationality or the other. To anticipate somewhat the argument of the next section, McMahon's PCR ultimately fails to ground a general obligation to obey the law. The PCR directs one to compare, from the standpoint of one's principles of value, the cooperative outcome with the outcome in which everyone defects, and hence, it can justify contribution (obedience to the law). On this view, however, whether one has sufficient reason to obey the law depends on the principles of value that one holds. Thus, even if the PCR is able justify a greater level of compliance with the law than the PIR in conjunction with an individual's values can justify, as long as we stick to McMahon's view according to which evaluative standards are essentially static, there is no escaping the conclusion that some individuals do not

have sufficient reason to comply with the law. Yet, admittedly, if a view according to which evaluative standards themselves can evolve were vindicated, this would make room for the possibility that everyone (or nearly everyone) acquires a reason to comply with the law.

4.6 The principle of collective rationality and the obligation to obey the law

Let me turn now to the issue of whether the PCR can ground a general obligation to obey the law. As already pointed out, in the political case, i.e. when the state is viewed as a cooperative scheme, contribution to the scheme takes the form of complying with the law. On McMahon's view, a member of a particular state has sufficient reason under the PCR to make her assigned contribution if she judges that life within the state is preferable to the best she could achieve if the state did not exist. The question is whether an appeal to the PCR can help meet the challenge raised by philosophical anarchists.

As McMahon himself emphasizes, there is no clear-cut answer to this question.[129] On the one hand, the PCR plainly dictates a greater level of compliance with the law than the PIR in conjunction with an individual's moral values dictate. We have seen above that individuals can have moral reasons for defecting (for breaking the law), and that as long as they consider the matter from the point of view of the PIR, there is no guarantee that the moral reason for contributing will outweigh the moral reasons for defecting. The problem of morally motivated defection is a central concern of philosophical anarchists, who insist that, while one can have a moral reason to obey a law, this reason will depend

[129] See McMahon, *Collective Rationality and Collective Reasoning*, pp. 65–66.

on the content of the law in question.[130] Philosophical anarchists have no quarrel with the claim that one has sufficient moral reason to obey a law when there is an independent moral reason to act as the law requires. However, they remain skeptical about the possibility of establishing that one can have sufficient moral reason to obey the law just because it is the law. Admittedly, the PCR can block the threat of morally motivated defection. What follows is that, unlike the substantive moral principles that work within the framework of the PIR, the PCR can be viewed as justifying obeying the law merely in virtue of its being the law.

The same conclusion can be reached from a different direction. McMahon claims that the PCR offers an illuminating account of legitimate preemption.[131] According to a prevalent view, political authority can be said to be legitimate to the extent that there is a sound basis for accepting the existence of a preemptive reason for complying with authority's directives. The PCR explains how de facto authority can be legitimate. McMahon emphasizes that:

> The essence of subordinating authority is preemption, an individual's deferral to a directive even when what it directs is something that he would not otherwise regard himself as having sufficient reason to do. […] the PCR can underwrite preemption because (1) it gives agents a reason to comply with a directive that they would not have reason to comply with if they considered the matter solely from the standpoint of the PIR, and (2) the PCR itself preempts the PIR. But to vindicate the preemption associated with the exercise of de facto political authority is – in one sense at least – to justify obeying the law as such, just because it is the law.[132]

[130] See Simmons, *Moral Principles and Political Obligations*, pp. 16–23, and *Justification and Legitimacy: Essays on Rights and Obligations*, p. 109.
[131] McMahon, *Collective Rationality and Collective Reasoning*, pp. 51–54.
[132] McMahon, *Collective Rationality and Collective Reasoning*, p. 66.

An elaborate account of how the PCR underwrites preemption must make reference to the fact that de facto subordinating authority provides a way of solving the assurance problem. We have seen that, according to McMahon, in cases when a group can benefit from cooperation, acceptance of the PIR by the members of the group would often lead to suboptimal outcomes. What is required in such cases in order to achieve the mutual benefits of cooperation is that the practical judgment authorized by the PCR takes precedence over the one authorized by the PIR. Yet, the PCR justifies contribution to a cooperative scheme only insofar as the assurance problem is solved (i.e. it gives each sufficient reason to contribute only insofar as there are reasons to believe that enough others will contribute so that the result will be an outcome that each can regard as preferable to the noncooperative outcome). McMahon argues that common knowledge within a group of the fact that a source of directives possesses de facto authority offers a solution to the assurance problem. On this line of thought, the directives of authority provide, under the PCR, a reason to act in particular way (a reason that the members of the group did not previously have) precisely because de facto subordinating authority solves the assurance problem.

So far we have seen that the PCR can be viewed as justifying obeying the law merely in virtue of its being the law. Yet, on the other hand, we should keep in mind that at the end of the day what one has sufficient reason to do under the PCR depends on the principles of value that one holds. As McMahon points out:

> The PCR expands our understanding of rational action in light of one's values, thus expanding the instances in which compliance with the law is

> justified, but the values of some individuals residing within the territory of a state could be such that compliance with the law remains unjustified.[133]

Before going further, let me briefly consider the question whether on the view under consideration one can have sufficient reason to comply with unjust laws. This is a particularly poignant question given that McMahon's account of what rationality requires in cooperative contexts is build around the assumption that individuals hold conflicting moral values. Let me detail.

McMahon calls attention to a point that often goes unnoticed in philosophical debates that focus on the question whether there is sufficient reason to comply with unjust laws. He points out that those who urge that one does not have to comply with morally objectionable laws advance a claim about what is required by morality, correctly understood. On this view, disobeying a law is morally permissible if the law in question is unjust. However, McMahon rightly emphasizes that this approach fails to take into consideration the problems posed by moral disagreement.[134] In other words, those who adopt this approach seem to lose sight of the fact that we cannot expect political cooperation to proceed under the assumption that a given conception of the good is the correct one, while others are mistaken. Nor can we, according to McMahon, expect cooperation to proceed under the assumption that a given conception of the right is correct. McMahon insists that political cooperation should be viewed as cooperation between individuals who hold different views about what justice requires. What follows from all these considerations is that maintaining a state is likely to require every one of its members to obey at least some apparently unjust laws.

[133] McMahon, *Collective Rationality and Collective Reasoning*, p. 66.
[134] McMahon, *Collective Rationality and Collective Reasoning*, pp. 70–71.

In short, McMahon maintains that, if we take evaluative diversity seriously, attempts to prove that upholding a state apparatus would only involve compliance with seemingly just laws are rendered highly implausible. This leads him to adopt the controversial view that one can have sufficient reason to obey unjust laws. Here is a summary of his argument.

In applying the PCR in order to determine whether one has sufficient reason to comply with a law, one can compare, from the standpoint of one's conception of justice, the existence of the legal order with the state of nature. Insofar as the legal order contains many laws that one judges to be unjust, one may find the state of nature to be preferable to the existence of the legal order. One may thus conclude that one has no reason to comply with the law as such. If, however, one judges that most laws are just and only a few are unjust, one will most likely find that the legal order is to be preferred to the state of nature. Thus, if to comply with a law means to contribute to the maintenance of the legal order, one may have according to the PCR sufficient reason to comply even with unjust laws. Another possibility is to ask whether, from the standpoint of one's conception of justice, the existence of a particular law which regulates a certain aspect of social cooperation is preferable to a situation in which no law regulates that aspect of cooperation. When the PCR is applied in this manner, one will most probably judge that one does not have sufficient reason to obey unjust laws. Surely, the question is whether individuals are more likely to apply the PCR to holistic or to piecemeal compliance. At this juncture, we must recall that, in proposing the PCR as the principle which governs the (first aspect of) rational cooperation, McMahon's primary interest is in offering a characterization of the dispositions of cooperatively disposed individuals. According to

him, individuals who are cooperatively disposed accept the PCR, and therefore, may judge that holistic compliance is to be preferred to the state of nature. Yet, McMahon contends that cooperatively disposed individuals would find holistic compliance to be preferable not only to the state of nature, but also to piecemeal compliance.[135] This is because, as long as each complies only with the laws that she considers just, implementing any controversial policy would be impossible, and thus some of the advantages that the existence of a legal order appears to have over the state of nature would be lost.

4.7 McMahon's argument reconsidered

Before moving on to the next chapter, let me consider a possible objection to McMahon's view that there is something deeply disturbing about suggesting that one has to manifest a disposition to make concessions when what is at stake is the realization of one's moral concerns. It will be useful to consider briefly a few remarks made by McMahon in his *Reasonable Disagreement: A Theory of Political Morality*. McMahon stresses that properly functioning individuals will reconsider the judgments with which others disagree.[136] In other words, properly functioning individuals take disagreement to provide a reason to reassess the case for their own position in light of what others hold. Note, however, that McMahon also points out that rational pressure to reconsider one's judgment in the face of disagreement "must be distinguished from rational pressure

[135] McMahon, *Collective Rationality and Collective Reasoning*, p. 69.
[136] McMahon, *Reasonable Disagreement: A Theory of Political Morality*, p. 69.

specifically to agree – to eliminate any disagreement that may remain after careful consideration".[137] So even though rationality requires, whenever competent reasoners disagree, a careful reconsideration of the arguments in favor of and against the view that is subject to disagreement, one may conclude that these arguments decisively favor one's own position on the matter. What is important to keep in mind that, on McMahon view, one cannot simply disregard the fact that others disagree. Disagreement provides a reason for thinking that one might be wrong, and thus, it prompts one to reassess the relevant arguments. Yet, McMahon also emphasizes that "it is not appropriate to treat this reason as evidence on a par with the relevant substantive reasons of which one might be aware", and that whether one should stick to one's view depends on such substantive reasons.[138]

A further consideration might prove relevant at this point. McMahon draws attention to an important consequence of the fact that properly functioning individuals reconsider the judgments with which others disagree. Properly functioning cooperators experience a rational pressure to reconsider their judgments in the face of disagreement, and thus, they present each other with arguments. So, in the political context, arguments for opposing views are publicly examined. Yet, according to McMahon, collective reasoning has the effect of eliminating, or at least reducing, incompetent reasoning. As he puts it, there is reason to expect that collective reasoning will eliminate unreasonable views.[139]

One might, however, wonder whether McMahon's contention that there is something dubious about requiring individuals to manifest a fundamental disposition to compromise when it comes to the realization of their moral concerns, loses some of its

[137] McMahon, *Reasonable Disagreement: A Theory of Political Morality*, p. 70.
[138] McMahon, *Reply to Gaus, Richardson and Weber*, p. 211.
[139] McMahon, *Reasonable Disagreement: A Theory of Political Morality*, p. 77.

initial plausibility once it is made clear that McMahon's concern is with conflicting, but nevertheless reasonable moral views. One might argue, for instance, that the initial plausibility of McMahon's claim derives at least partly from the fact that, when we consider the question of whether showing a disposition to compromise when it comes to one's moral concerns is appropriate, we tend to think about the perils that any yielding to unreasonable moral views would involve. However, the argument would go, it is not entirely clear that we would just as readily endorse the claim that there is something dubious about requiring individuals to manifest a disposition to compromise when it comes to their moral views, once it becomes apparent that we are asked to consider only contexts in which moral conflicts involve incompatible, but nevertheless reasonable moral views.

It should also be stressed here that, in addressing the issue of how to characterize the reason to contribute to cooperative schemes, McMahon explicitly states that his aim is to offer a characterization of the dispositions of cooperatively disposed individuals.[140] In other words, instead of being concerned with establishing that contributing to cooperative enterprises is rational even when one could do better by free riding, he starts from the observation that many individuals do in fact contribute to cooperative schemes and that, in doing so, they feel confident that they act in accordance with the requirements of rationality. Employing the method of wide reflective equilibrium, McMahon claims that this is reason enough to accept that contribution to mutually beneficial cooperative enterprises is justified, and that the theoretical task at hand is simply to show how the requirement to contribute is best understood. What interests me here is that, in raising the issue of whether the PCR offers a better characterization of this requirement than the

[140] McMahon, *Collective Rationality and Collective Reasoning*, pp. 15, 27.

principle of fairness, McMahon envisages a group of cooperatively disposed individuals. Yet, we have already seen above that properly functioning cooperators typically engage in collective reasoning about practical matters, and that collective reasoning has effect of eliminating unreasonable moral views. So, McMahon's argument to the effect that the PCR is preferable to fairness as a characterization of the reason to contribute to cooperative enterprises because to invoke fairness in adjudicating between conflicting moral points of view would mean to demote moral concerns to interests, should be understood as applying in a context in which any two competing moral views will, in fact, be equally reasonable.

However, a proponent of McMahon's view might insist that his argument does not depend on whether the competing moral views are reasonable. Consider the following excerpt from McMahon:

> For the person holding a moral concern […] its importance is not exhausted by the fact that satisfying it satisfies her. It is taken to identify a feature that the world morally ought to have. This is compatible with accepting situations in which one's moral concerns are not as fully realized as one would like them to be. One does not demote one's moral concerns to interests by doing the best one can in an unfavorable situation. But acquiescing in the less than maximal realization of one's moral values because *fairness* requires this does alter the character of one's moral concerns. If one has the power to bring about a morally preferably alternative, but refrains for reasons of fairness, one treats one's concerns as interests. And it appears that the demotion of moral concerns to interests is something to regret.[141]

Still, one can point out that McMahon's argument relies on an equivocation between different types of situations in which considerations of fairness might be invoked. There are several cases that we need to take into consideration. Consider, first,

[141] McMahon, *Collective Rationality and Collective Reasoning*, p. 95 (emphasis in original).

cases that do not involve a conflict between two moral views. Assume, however, that the individual may use the resources that he is called upon to contribute in order to promote his own moral view. Admittedly, McMahon would agree that, in this case, invoking considerations of fairness is not inappropriate. It should be noted that, when discussing the issue of the choice of a cooperative scheme, McMahon stresses that fairness requires equal opportunity for political influence.[142] Given this emphasis on the idea of equal opportunity to promote one's conception of the good, it is reasonable to infer that McMahon would agree that, in this case, using the resources that one is called upon to contribute in order to promote one's own moral view is appropriately condemned by the principle of fairness.

The second type of situations that we need to take into considerations is that in which, as McMahon suggests, "one has the power to bring about a morally preferably alternative, but refrains for reasons of fairness". This might be due to a worry about whether everyone's conception of the good is realized to the same extent.[143] In this case, we can agree with McMahon that "acquiescing in the less than maximal realization of one's moral values because *fairness* requires this'' amounts to demoting moral concerns to interests.

Finally, consider a case which involves a conflict of moral views. Suppose that this is a case in which not all members of the group regard the choice of the cooperative scheme as appropriate. Suppose further that a member of the minority which does not view the choice of the scheme as justified uses the resources that he is called upon to contribute in order to promote his own moral view. This is a case in which considerations

[142] McMahon, *Collective Rationality and Collective Reasoning*, pp. 94–96.
[143] For arguing against such a meta-fairness doctrine, see Gaus, pp. *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World* 406–410.

of fairness are supposed to arbitrate between moral views, and McMahon might be right to hold that it is not entirely clear that considerations of fairness should prevail in this case.

To sum up, while we have reason to agree with McMahon that there is something deeply troubling about suggesting that one has to manifest a fundamental disposition to make concessions when what is at stake is the realization of one's moral concerns, there are, nevertheless, several considerations that can brought in defense of the idea that considerations of fairness can appropriately be invoked when attempting to justify political obligation. First, the plausibility of McMahon's claim has to do with the fact that, whenever we think about whether refraining from promoting one's moral concerns is appropriate, we tend to think about the perils that any yielding to unreasonable moral views would involve. Second, even if we grant that refraining from promoting one's moral concerns on the basis of considerations of fairness is inappropriate under certain circumstances (the second case considered above), this has no bearing on the question of whether such considerations can legitimately be invoked when attempting to justify political obligation. Finally, even if we focus on circumstances that are relevant for the problem of political obligation (the third case considered above), McMahon's argument to the effect that fairness cannot justify political obligation depends on the claim that it is not possible that the choice of a cooperative scheme is regarded by all as guided by reason.

Chapter 5
Normative conventionalism

5.1 Introduction

To look at the merits of contemporary conventionalism and to asses the claim that, as a theoretical framework, conventionalism has valuable resources to provide an innovative insight into both the problem of justifying legal and political authority and the problem of political obligation, I will begin in the next section by offering a rough picture of the way in which conventionalism enters contemporary debates. In the following sections, I proceed by looking at the arguments put forward by David Lewis in support of his own version of conventionalism. I then proceed by looking at the main claims that should characterize a plausible conventionalist account of authority. According to the conventionalist account of authority, if a rule is legitimate, this constitutes a reason for following it. Yet, on this account, a rule can only be legitimate if it is part of a system that is by and large effective. In other words, a rule's being legitimate comes down to its being part of a system of rules that are generally followed. However, authors like Eerik Lagerspetz argue that conventionalism can provide a way out of this apparently vicious circularity.[144] Arguably, this circularity merely reflects the interdependence of individuals' attitudes towards authority.

The conventionalist account that I focus on takes the "interdependence of reasons" as the basic concept. According to this account, individuals have reasons to act in a certain way only because they believe that others have reasons to act in the same

[144] Eerik Lagerspetz, *A Conventionalist Theory of Institutions*, p. 78.

way. This begins to explain the interest of those who advocate conventionalist accounts of authority in game theory. In looking for situations that make intelligible the interdependence of individuals' reasons for action, one is naturally led to consider the sort of games that game theory deals with, in which the outcome for each player depends not only on his own choices, but also on the choices made by the other players. While one can argue that a game theoretical approach enhances the plausibility of the idea that, if there are reasons to follow the directives of authority, such reasons can only be interdependent ones, a common problem seems to confront the proponents of such approaches. As Govert den Hartogh rightly points out, while in the case of coordination games, the players have reason "to make their own choices fit into the pattern of interdependent decisions", in other cases, e.g. in prisoner's dilemma situations, the players seem to have "reason rather to deviate from the pattern – which makes it difficult to understand how the pattern can either emerge at all, or be sustained once it has emerged".[145] Yet, given that social interaction regulated by law is not restricted to coordination games, but consists also of games such as prisoner's dilemmas, conventionalists who appeal to game theory would have to account for the fact that in the latter type of cases individuals could also have reason to make their choices fit into a pattern of interdependent choices. One solution is to invoke, as den Hartogh proposes, the exercise of certain cooperative virtues. One of my aims in this chapter is to assess the merits of this proposal. I will attempt to do this by trying to determine at what level it is plausible to claim that cooperative virtues enter the picture. Recall from Chapter 4 that McMahon argues that the thesis that people are cooperatively disposed functions within his theory as a presupposition. The considerations presented in this chapter lead me

[145] Govert den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, pp. 6–7.

towards a qualified acceptance of this presupposition, and towards an analysis of a further claim, namely that a proper characterization of political obligations will necessarily include an appeal to the same presupposition – that people are indeed cooperatively disposed. In this chapter, I argue for the plausibility of an understanding of political obligation developed within a conventionalist framework qualified in such a way as to make room for the normative use of substantive moral principles such as the principle of fairness.

5.2 Preliminary remarks on conventionalism as a normative theory

Let me begin by considering the relationships that hold between conventionalism, contractarianism and consent theories. One reason for taking up such a task is the need to establish whether any one of the three theoretical approaches is better suited to answer a certain philosophical question (e.g. 'are there any grounds for suitably general political obligations?'). Another such reason is the attempt to establish a certain "framework" that details the way in which these three theoretical positions can interact, with the hope that this framework itself would eventually constitute a theoretical advance which, in turn, could shed new light and advance the thinking on a number of topics.

Attempting a similar analysis, Leslie Green notices that conventionalism should, in fact, be contrasted with both contractarianism and consent theories. According to him, conventionalism and contractarianism as contrasted with consent theories, differ in that the first two are "fundamentally a theory of social order, while the latter is not; it is a

theory of social relations".[146] Still, Green claims that no theoretical advantage can come out of keeping them together, as traditional political argument does and, furthermore, that each of these theories must be assessed on its own merit. Conventionalism and contractarianism are further different, Green says, on several levels. First, with respect to their motivational structure: while both are "ways of creating social order", conventions attempt to remedy "problems of information among parties sharing a common interest", while contracts attempt to remedy "problems of motivation among parties with partly common and partly conflicting interests".[147] Second, conventions "need not be *enforced* in order to be *in force*": they must only be stable, so that for no individuals it pays to 'go it alone', but they do not need to be optimal, nor do they need to maximize utility.[148] Conventions differ from other regularities that provide mutual benefit in that they lack the added temptation for free-riding that these ones retain. In this sense, conventions are only "those situations of coordination where each prefers matching his own behaviour with that of the others to not matching at all, and is not so attracted to attracted to any particular way of matching that he is ever tempted to go it alone". Hence, Green believes that "problems of collective action and public goods are not the correct baseline or state of nature with respect to which conventional norms are to be analysed. Instead, they provide the logical model for contractarian arguments".[149]

While not being a conventionalist himself, Green admits that conventions constitute an interesting venue for the study of political authority, both historically (because they represent an argument often invoked) and theoretically (because, since they

[146] Leslie Green, *The Authority of the State*, p. 94.
[147] Leslie Green, *The Authority of the State*, p. 93.
[148] Leslie Green, *The Authority of the State*, p. 93.
[149] Leslie Green, *The Authority of the State*, p. 93.

lack the problematic aspect of enforcement, they make the search for an indirect justification of authority a lot easier). In historical terms, among the chief advantages that conventions possess is the fact that they are immune to the Humean attack on consent theories (i.e. almost none of the citizens of a *polis* give their consent in the way required by the theory, hence any concept of consent useful for theoretical purposes is too strong and impractical for practical purposes, etc). To work, conventions are in no need of convening parties and agreements. In fact, under specified background conditions, all they need is rational behaviour. As Green puts it, "even if 'tacit consent' is not possible, tacit conventions are".[150] There is, however, an important difference between the two approaches. Conventional theories are always instrumental. Conventions can only be means of promoting a common interest that already exists, independently of anything that conventions (presuppose or) bring into play. This is not always so with consent theories, for consenting to obey can be, as Green notices, "an expression constitutive of a certain relationship", and not merely the means for bringing about certain benefits or goods, of for furthering certain interests.[151]

5.3 David Lewis on convention

Game theory is just one of the areas of philosophy in which David Lewis has systematically demonstrated his distinctive philosophical prowess. He is commonly

[150] Leslie Green, *The Authority of the State*, p. 94.
[151] Leslie Green, *The Authority of the State*, p. 94.

credited[152] with two major contributions to the field: the introduction of the concept of common knowledge, and the first thorough analysis of convention, the latter constituting, undoubtedly, the most widely discussed, thoroughly criticized, and massively employed account of convention to date.[153] In the present discussion, I will constantly return to Lewis's account of convention and I will also refer, along the way, to several points he makes on issues related to conventions. This being so, it is preferable to offer a detailed examination of his views on these topics at the outset. Examining Lewis's account is essential for my purposes not only in virtue of its being the most thoroughly discussed and widely accepted contemporary account of convention, but also because several of its characteristics make it very important for my general outlook.

Firstly, Lewis offers an individualistic account of conventions, i.e. one that deals with conventions exclusively in terms of individual agents and their beliefs, preferences and actions, as opposed to corporatist (or collectivist) accounts of conventions, which are structured in terms of irreducible collective entities (instead of individual agents) or of *sui generis* social practices (instead of individual actions, intentions, beliefs, preferences and attitudes). Secondly, because it seeks to provide good reasons for an agent to conform to a convention, Lewis's account is also a rationalist account, standing, in this respect, in opposition with markedly non-rationalist accounts such as those put forward, among

[152] For this point, see Robert Cubitt and Robert Sugden *Common Knowledge, Salience and Convention: A Reconstruction of David Lewis' Game Theory*, p. 175.

[153] For Lewis's initial account, see his *Convention: A Philosophical Study*; for a version which includes certain revisions, see his *Language and Languages*. A sample of authors who criticize Lewis includes Tyler Burge (in *On Knowledge and Convention*), Margaret Gilbert (in *On Social Facts* and *Rationality and Salience*), and Simon Blackburn (in *Spreading the Word*). Among the authors who employ one version or another of the Lewisian analysis of conventions in their works are Jules Coleman, Robert Cubbitt, Govert den Hartogh Julie Dickson, John Elster, Joseph Kutz, Eerik Lagerspetz, Brian Leiter, Andrei Marmor, Seumas Miller, Gerald Postema, Scott J. Shapiro, Robert Sugden, Peter Vanderschraaf and Bruno Verbeek.

others, by Winch[154] and Gilbert[155], according to which no *rational* justification of an agent's conformity to a convention is possible or, indeed, necessary.

Finally, Lewis's account is important for yet another reason. Bearing in mind the distinction between two major ways of understanding social interactions - one that goes along the lines of coordination games and conventions (a move that results in the study of *coordination*), and another one that goes along the lines of 'Prisoner's Dilemma games' (a move that results in the study of *cooperation*), it is important to notice that Lewis's account seemingly deals only with the first of the two ways. If this would be so, it would appear that an extensive treatment of his views is unwarranted. In fact, as it will soon become obvious, Lewis does more than that. He offers an insight into two general classes of social interactions which can be analyzed using an account of convention: one class,

---

For the two distinctions (i.e. individualistic vs. corporatist/collectivist and rationalist vs. non-rationalist accounts), see Seumas Miller, *On Convention*, p. 435. For Winch's account, see Peter Winch, *The Idea of a Social Science and Its Relation to Philosophy*, p. 30ff. Winch, in fact, interprets Wittgenstein's analysis of what it is *to follow a rule* in order to establish that "a very important feature of the concept of following a rule… [is] that one has to take account not only of the actions of the person whose behavior is in question as a candidate for the category of rule-following, but also the reactions of other people to what he does. More specifically, it is only in a situation in which it makes sense to suppose that somebody else could in principle discover the rule which I am following that I can intelligibly be said to follow a rule at all." (p. 30). Attempting to avoid possible misunderstandings, he follows through by admitting that it is possible, "within a human society as we know it, with its established language and institutions, for an individual to adhere to a *private* rule of conduct", but concludes, on what I take to be, in the abovementioned sense, a non-rationalistic tone, by stressing a most important lesson from Wittgenstein, namely that "it makes no sense to suppose anyone capable of establishing a purely personal standard of behaviour *if* he had never had any experience of human society with its socially established rules" (p. 33, Winch's italics). Since, as I just said, his is mainly an argument about what it is to follow a rule, and not about what it is to act in conformity with a convention, my claim should be taken in the weaker sense that, by insisting in more general terms, other than those of the analysis of rule-following, that it makes no sense to start from the presupposition that standards of behavior can be individually created and, hence, that such activities are always socially-dependent (and not the result of the preferences, beliefs and attitudes of individual agents), Winch adopts a stance which seems both non-rationalist and, as a matter of fact, anti-individualist as well.

[155] In *On Social Facts* (Chapter VI, especially pp. 318ff), Gilbert discusses Lewis's views and criticizes them on multiple accounts. While I will revisit some of Gilbert's points later on, it is worth mentioning now that some of the most important charges that she levels against Lewis are directed against the traits that give his approach its rationalist character. For example, Gilbert claims that "it is not true that rational agents will inevitably come to follow a successful precedent in a particular coordination problem, other things being equal" (p. 330), and that it should be acknowledged that "rational agents as such cannot be expected to do their part in a given salient solution in an otherwise problematic coordination problem", because otherwise there is a risk that "we take the exercise of rationality to be a more useful tool than it is" (p. 336).

that of problems of coordination, consists of situations in which there is, among all the agents involved, a perfect or nearly perfect coincidence of interests; another class, that of problems of partial conflict, consists of situations in which the gain of some of the agents involved is only possible if some of the other agents make certain sacrifices.[156] It seems then that Lewis's overall approach is broad and encompassing enough to be granted a separate analysis, even without a recurrent reference to issues engendered by the game theoretical study of cooperation.

Let me add one final clarification, by briefly looking at what a further Lewisian distinction implies for the way I will proceed. According to Lewis, conventions should primarily be analyzed as solutions to recurring coordination problems. A distinction needs to be made, he points out, among regularities that hold within a group, but can be fully explained by historical agreements, and regularities which are, within the same group, properly speaking conventional. Among the latter, the focus is on regularities that are beneficial for every member of the group. As it turns out, such regularities will appear when the group is faced with a coordination problem, when it is best for each member of the group if they all follow the same regularity and they have an interest in coordinating themselves so that they, in fact, do so. Now, Lewis offers two different definitions of coordination problems - one of them, more technical, couched in game theoretical terms, the other in terms of conditional preferences. I will consider each of them in its turn shortly, i.e. in the next sections of the chapter. Before that, however, a preliminary look at some distinctions that concern coordination problems and an attempt to spell out the

[156] For this distinction, see Peter Vanderschraaf, *Knowledge, Equilibrium and Convention*, p. 338.

reasons why game theorists, political philosophers and legal theorists alike are so interested in studying such problems seem necessary.[157]

5.4 Conventions as solutions to coordination problems

In somewhat general terms, a coordination problem is a situation in which the interests of the parties coincide in that, when faced with a choice among a set of alternative actions, each of the parties will rank higher and will, eventually, opt for that particular action which will likely be done by most of the parties involved. It is not necessary for all the agents to have the same set of alternative actions. Besides, since the outcomes the agents want to produce or prevent are, as Lewis says, "determined jointly by the actions of all the agents", and the outcome of any action an agent might choose will depend on the actions of the other agents, it will make no difference from which of the possible sets of alternative actions the chosen action is derived. What is important is to note that "each must choose what to do according to his expectations about what the others will do".[158]

So described, coordination problems are obviously interesting, because they seem to provide a way of deriving normative models for the actual structure of an important

[157] See Leslie Green (in *The Authority of the State* and several of his articles), Joseph Raz (in *Practical Reason and Norms*, *The Morality of Freedom* and several of his articles; at one point, Raz wrote that "Solving coordination problems is one of the important tasks of political and many other practical authorities" - *The Morality of Freedom*, p. 31), Andrei Marmor, Jules Coleman, Gerald Postema, Jeremy Waldron, Owen M. Fiss (in *Conventionalism*, especially pp. 177–197), E. Weinrib, Th. Benditt, Govert den Hartogh, Eerik Lagerspetz, among legal philosophers, Thomas Schelling, David Lewis, Robert Aumann, Jon Elster, Robert Sugden, Peter Vanderschraaf among game theorists, and David Gauthier, Jean Hampton, Gregory Kavka, Brian Barry, Christopher Morris among political philosophers.

[158] David Lewis, *Convention: A Philosophical Study*, p. 8.

number of social interactions. Norms plausibly arrived at by studying coordination problems range from the often mentioned decision about the side on which everyone needs to drive (when being in the position of having to settle the issue, it is important to notice that it does not really matter on which side one drives, as long as everybody else drives on the same side), to money (difficult barter processes are avoided by settling on a common monetary currency), and from norms regarding the adoption or establishment of a common language to various other norms regulating those circumstances in which a thorough reciprocal knowledge of the preferences of the relevant parties is necessary. In all situations in which the members of a group that need to interact will find themselves in need to coordinate in order to find a solution out of what is, initially, a neutral, yet problematic, context of choice among available actions.[159]

5.4.1 Lewis on coordination problems: a game-theoretical reading

In discussing coordination problems, Lewis takes as reference point a classification put forward by Thomas Schelling in his seminal book *The Strategy of Conflict*. According to Schelling's proposed "reorientation of game theory", games, as problems of interdependent decision, can be represented as positioned across a spectrum ranging from *games of pure conflict* to *games of pure coordination*.[160] Games of pure

[159] Apart from deciding on which side to drive, having to agree on a common language and settling upon a common currency, Lewis also offers other examples which are relevant for his approach of coordination. Such are those of two persons wanting to meet each other, of two persons who are cut off while talking on the telephone, of oligopolists attempting to establish a unique price, or Hume's example of two men rowing a boat. Later on, I will briefly revisit some of these other examples.

[160] Thomas Schelling, *The Strategy of Conflict*, pp. 83–118, 291–303.

conflict are, in Lewis's terms, those in which "the agents' interests are perfectly opposed". Such games will be represented by a payoff matrix in which the sum of the agents' payoffs is zero in every square, and it will always be the case that "one agent's losses are the others' gains, and vice versa". Games of pure coordination, on the other hand, are games in which "the agents' interests coincide perfectly". Such games can be represented by a payoff matrix in which the sum of the agents' payoffs is equal in every square.[161]

	C 1	**C 2**	**C 3**
R 1	0 0	-.5 .5	-.5 .5
R 2	.5 -.5	1 -1	-1 1
R3	.5 -.5	-1 1	1 -1

Fig. 1 Payoff matrix for a game of pure conflict

Needless to say, equilibria will exist in games of pure coordination. This, however, does not imply that no equilibria can exist in a pure conflict game. In the example from figure 1 above, **<R1, C1>** is an equilibrium: Row prefers it to both **<R2, C1>** and **<R3, C1>**, and Column prefers it to both **<R1, C2>** and **<R1, C3>**. It is important to be unambiguous about this point, and being so allows Lewis to stave off a possible objection to his views. As he remarks, those coordination problems in which he is mostly interested are "among the situations at or near the pure coordination end of

[161] David Lewis, *Convention: A Philosophical Study*, pp. 13–14.

Schelling's spectrum".[162] So, it could be objected, if Lewis would assume by definition that coordination games are characterized by the presence of equilibria, this would not only go, illegitimately, against the grain of game theory, but also load the dice, equally illegitimately, in favor of his definition. Lewis notices that the opposite is actually true, and concludes that it is improper to claim that coordination problems are uniquely distinguished by the presence of equilibria.[163]

What, then, does distinguish coordination problems, according to Lewis's treatment, apart from their being at or near the pure coordination end of Schelling's spectrum? Two aspects stand out. The first one is that, since coordination problems are positioned merely "near" the end of this spectrum (and not right at its end), it is possible that *impure* coordination problems exist, i.e. that coordination problems exist which allow for the interests of the players to be partially conflicting. One of Lewis's examples clearly illustrates how allowing for *imperfect* coincidence of interests works, according to his views. He writes:

> Suppose you and I both want to meet each other. We will meet if and only if we go to the same place. It matters little to either of us where (within limits) he goes if he meets the other there; and it matters little to either of us where he goes if he fails to meet the other there. We must each choose where to go. The best place for me to go is the place where you will go, so I try to figure out where you will go and to go there myself. You do the same. Each chooses according to his expectation of the other's choice. If either succeeds, so does the other; the outcome is one we both desired.[164]

[162] David Lewis, *Convention: A Philosophical Study*, p. 14. All other games will be positioned in Schelling's spectrum in accordance with the proportions of opposition and coincidence of interests, of conflict and coordination that they include.
[163] He writes: "Indeed the bulk of mathematical theory of games is precisely the theory of equilibrium combinations (known as *saddle points* or *solutions*) in situations of the opposite kind: pure conflict of interests between two agents" (p. 13).
[164] David Lewis, *Convention: A Philosophical Study*, p. 5.

When considering this example in order to characterize coordination problems, it can be allowed that each of the players cares, to a certain extent, about where they will end up going, although they care about this much less than about whether they meet. One of the possible payoff matrices which illustrate the fact that imperfect coincidence of interests is allowed is the one in Figure 2 below. While **<R1, C1>, <R2, C2>** and **<R3, C3>** are all equilibrium outcomes, they are not indifferent to the players: both Row and Column prefer **<R1, C1>** to the other two, but it is also true that they both prefer any of the three outcomes to any of the nonequilibrium outcomes.[165]

	C 1	**C 2**	**C 3**
R 1	1.5 Meet 1.5	.2 .5	0 .5
R 2	.5 .2	1.2 Meet 1.2	0 .2
R3	.5 0	.2 0	1 Meet 1

Fig. 2 Payoff matrix for 'the meeting point example' in which imperfect coincidence of interests is allowed

The second aspect which is characteristic of Lewis's treatment of coordination problems is even more important. It has to do with Lewis's focus on analyzing those situations in which coincidence of interests is predominant. These are cases in which, like in the matrix in Figure 2, "[…] the differences between different agents' payoffs in any one square (perhaps after suitable linear rescaling) are small compared to some of the

[165] David Lewis, *Convention: A Philosophical Study*, p. 10 and pp. 14ff.

differences between payoffs in different squares".[166] Focusing on such cases allows him to give a definition to the notion of coordination equilibrium, which is central to his treatment of conventions.

Equilibria in general are, according to Lewis, combinations of actions in which each agent has done as well as he can given the actions of the other agents. Within an equilibrium combination, no agent could have produced a better outcome for himself, given the actions of the other agents. Two observations are in place here. Firstly, note that no implication about an equilibrium being optimal for the agent follows because, as Lewis says, even though all combinations of actions that exist which are best for everyone are equilibria, this does not mean that "an equilibrium combination must produce an outcome that is best for even one of the agents. […] In an equilibrium, it is entirely possible that some or all the agents would have been better off if some or all had acted differently. What is not possible is that any one of the agents would have been better off if he alone had acted differently and the rest had acted just as they did".[167] Secondly, note that Lewis rephrases the abovementioned definition of an equilibrium outcome as a combination which "[…] each agent likes *at least as well as* any other combination he could have reached, given the others' choices." From this way of rephrasing the original definition Lewis derives the definition of a proper equilibrium,

[166] David Lewis, *Convention: A Philosophical Study*, p. 14.

[167] David Lewis, *Convention: A Philosophical Study*, p. 8. A huge literature on equilibrium theory exists, and an attempt to review it here could not accomplish much. Let me note, however, two highly relevant contributions. One definition which is similar to the one advanced by Lewis is given by James Friedman, who writes that the condition which applies to all equilibrium outcomes is that: "No single player would have obtained a larger payoff had she used an alternative strategy, given the strategies of the other players" (James Friedman, *Game Theory with Applications to Economics*, p. 3). David Kreps adds a supplementary condition to the characterization of equilibria which is relevant for my discussion, because it underscores the importance of expectations (or of "conjectures", as Kreps prefers to call them) for the stability of equilibria. According to him, "The equilibrium actions of the [players] are consistent with the conjectures that each [player] is supposed to hold" (David Kreps, *A Course in Microeconomic Theory*, p. 330).

according to which a combination is "[…] a *proper* equilibrium if each agent likes it *better than* any other combination he could have reached, given the others' choices."[168]

Within this framework, in order to proceed to a characterization of conventions in terms of coordination equilibria, one must first define the notion of coordination equilibrium. On this point, Lewis writes:

> Let me define a *coordination equilibrium* as a combination in which no one would have been better off had *any one* agent alone acted otherwise, either himself or someone else. Coordination equilibria are equilibria, by the definitions. Equilibria in games of pure coordination are always coordination equilibria, since the agents' interests coincide perfectly. Any game of pure coordination has at least one coordination equilibrium, since it has at least one outcome that is best for all. But coordination equilibria are by no means confined to games of pure coordination. They are common in situations with mixed opposition and conflict of interests. They can occur even in games of pure conflict […] Most versions of our sample coordination problems are not games of pure coordination; but they all have coordination equilibria.[169]

Thus, we are finally in possession of the necessary elements needed to spell out Lewis's understanding of coordination problems couched in game-theoretical terms. According to it, coordination problems "are situations of interdependent decision by two or more agents in which coincidence of interests predominates and in which there are two or more proper coordination equilibria."[170]

[168] David Lewis, *Convention: A Philosophical Study*, p. 22.
[169] David Lewis, *Convention: A Philosophical Study*, pp. 14-15.
[170] David Lewis, *Convention: A Philosophical Study*, p. 24.

5.4.2 Lewis on coordination problems: a conditional preferences reading

Let me point out from the outset that the notion of conditional preference is intended to cover preferences for actions. In this sense, one agent's conditional preference for performing (or refraining from performing) an action is a preference the agent would have if and only if the other agents would themselves perform (or refrain from performing) particular actions. Let us call coordination problems defined in terms of conditional preferences in this manner formal coordination problems, and distinguish them from informal coordination problems, which is are simply problems that appear from having to choose between two courses of action which are more or less equally feasible but mutually exclusively. Understood in this way, formal coordination problems are intended to sum up in more exact terms the nature of informal coordination problems.[171]

Also at the outset, let me bring in Lewis's own definition of convention in terms of conditional preferences:

> "A regularity *R* in the behaviour of members of a population *P* when they are agents in an recurrent situation *S* is a *convention* if and only if it is true that, and it is common knowledge in *P* that in almost any instance of *S* among members of *P*,
>
> 1. almost everyone conforms to *R*;
> 2. almost everyone expects almost everyone else to conform to *R*;

[171] For this distinction, see Seumas Miller, *Social Action: A Teleological Account*, p. 94 ff.

3. almost everyone has approximately the same preferences regarding all possible combinations of actions;

4. almost everyone prefers that any one more conform to *R*, on the condition that almost everyone conform to *R*;

5. almost everyone would prefer that any one more conform to *R′*, on the condition that almost everyone conform to *R′*,

where *R′* is some possible regularity in the behaviour of members of *P* in *S*, such that almost no one in almost any instance of *S*, among members of *P* could conform both to *R′* and to *R*."[172]

Within this framework, it becomes clear how the two readings of conventions are connected. As it stands, the second understanding of what a convention is, given in terms of conditional preferences, cannot be only about looking at the actions which agents who possess a given set of conditional preferences perform in a particular recurrent situation. Because, per the definition above, each agent has both knowledge of and expectations about the actions of the other agents, this understanding of convention must be taken as also offering a suggestion about the way in which its central notion, that of conditional preference, is related to the central notion of the first understanding of convention, i.e. the notion of a conventional alternative. It seems plausible to argue that Lewis's claim that it is part of the very definition of a convention that a conventional alternative to it exists closely parallels the idea of conditional preferences. As it has already been observed,[173]

[172] David Lewis, *Convention: A Philosophical Study*, p. 78.
[173] By Seumas Miller, *Social Action: A Teleological Account*, p. 95.

the agents do not have unconditional, but merely conditional preferences for the conformity with a regularity *R* precisely because a conventional alternative *R′* is available.

5.5 Conventionalist accounts of authority

In *The Authority of the State*, Green puts forward a series of arguments against the conventionalist theory of authority. His arguments focus mainly on the theories of Joseph Raz and John Finnis, both of whom develop, as it has been noticed, "a conventionalist theory of authority (in a wide sense of the term)", but he treats their conventionalism via the more technical understanding of convention that was put forward by David Lewis, Edna Ullmann-Margalit, and Gerald Postema.[174] This being so, one can argue that, concerning precisely their main aim, Green's arguments are misguided, since both Raz and Finnis moved away, in their later work, from the technical treatment of conventions and coordination problems offered by game theory.[175] However, putting some distance between one's favorite approach and an already existing one is only sometimes necessary; some other times, it is not. For instance, if attempting to distance one's

[174] Lagerspetz, *The Opposite Mirrors: An Essay on the Conventionalist Theory of Institutions*, p. 93. For this point, see also Green, *The Authority of the State*, p. 111: "At least on the issue of law's co-ordinative function and its role in the justification of political authority, Raz the positivist and Finnis the natural lawyer seem to be in broad agreement". For putting forward a technical understanding of conventions, see David Lewis, *Convention: A Philosophical Study*, Edna Ullmann-Margalit, *The Emergence of Norms*, and Gerald Postema, *Coordination and Convention at the Foundations of Law*.

[175] For this point, see Lagerspetz, *The Opposite Mirrors: An Essay on the Conventionalist Theory of Institutions*, p. 93, bringing Finnis's *The Authority of Law in the Predicament of Contemporary Social Theory* and Raz's *Hart on Moral Rights and Legal Duties*, as well as his personal communication with both Raz and Finnis in support of his claim. My discussion of Green's arguments draws on Lagerspetz's own discussion at this and at several other, related points (see below). For Lagerspetz's discussion, see *The Opposite Mirrors: An Essay on the Conventionalist Theory of Institutions*, especially pp. 93–104.

treatment of the role of coordination in the study of authority from the model offered by game theory has to do, say, with a concern about the implausibility of importing certain unwarranted, theoretically undesirable or unnecessary assumptions characteristic of game theory in general, then it can only be said that such a concern is, in itself, unwarranted. While I realize that my claim stands in need of clarification, I believe that the aim to keep this discussion focused is better served if, at this point, I will eschew the task of providing the needed clarification. Taking up this task would necessitate, among other explanatory considerations, a longer explanation of what is and what is not assumed by game theory, and this move would change entirely the direction of this section.[176] Instead, let me just note for now that Lewis's treatment of conventions and coordination problems can largely be reconstructed without any or with a minimal use of game-theoretical terms, which suggests that, for him, using the framework of game theory is primarily an illustrative enterprise. Still, even if this is true, it does not say much about what reasons are there for giving up an approach couched in game-theoretical terms (other than ones having to do, perhaps, with a certain simplicity of presentation), and it certainly does not say anything about the legitimacy of applying the framework of game theory to the study of human interactions in general, and to the relationship between coordination and authority, in particular. For now, however, it is important only to stress that a) it is *legitimate* to study issues such as the relationship between authority and coordination within the framework of game theory, and that b) it is an open argument whether a study

[176] Many good discussions of the assumptions employed by game theory exist in the literature. For a good and brief such discussion, see Lagerspetz, *The Opposite Mirrors: An Essay on the Conventionalist Theory of Institutions*, esp. pp. 30–33. I will have more to say both about the role of certain assumptions of game theory, and about the way in which they enter the study of coordination problems and of authority later on.

conducted on these terms is worthwhile, i.e. whether it provides a substantive insight and advances the topic.

These points are all the more important when discussing the arguments put forward by Leslie Green, since his views go directly against this approach. According to Green, in dealing with coordination problems, appeals to authority are not necessary. What is more, such appeals must be rejected even when they might be sufficient to solve a coordination problem because, by comparison with other possible solutions, they possess no "evolutionary advantage". While Green does not substantiate the notion of evolutionary advantage any further, it seems that what he means by it is simply rationality, and this becomes clear once we look at the listed disadvantages that appealing to authoritative solutions implies.[177] If this is so, then his claim is that, in fact, an alternative rational solution that needs no authoritative decision is available in any coordination equilibria. Examples of alternatives to an appeal to authority or, as Green refers to them, non-authoritative methods of decision-making, range from coercion and bargaining (or negotiation) to self-coordination (or ad hoc convention formation). They can always be employed for reaching equilibrium in coordination situations. Green writes that:

> To achieve a co-ordination equilibrium by appealing to or creating a conventional norm, one need only act on the balance of first order reasons. Even in the absence of existing facts, one can solve a CP just by providing ordinary reasons in favour of one of the alternatives. One could suggest, advise, persuade or threaten the others to follow it. Each would act rationally in weighing the likelihood of the rest following that option, O_1, as opposed to any $O_2,\ldots,O_n$. When there is no existing reason to choose any of them over the others, and the first-order reason is then given for choosing O_1, it follows that there is some reason to choose O_1 and no reason to choose $O_2,\ldots,O_n$. If all other options are outweighed, then O_1

[177] Leslie Green, *The Authority of the State*, pp. 114–115.

> will enjoy general compliance. Hence, equilibrium can be secured without authority.[178]

The thrust of this argument is that authoritative decisions cannot function as exclusionary reasons simply due to the fact that, in coordination situations, no excluded reasons could exist, since the parties involved face, by definition, an indifferent choice between the available paths to reach the equilibrium:

> A conventionalist might, however, further object that while the existence of a CP always provides some reason for conformity it may not be an overriding reason. Each may still have preferences among the possible outcomes such that authoritative requirements are needed to single out a particular one. But this plays fast and loose with the description of the problem. If contrary preferences are strong enough to outweigh preferences for conformity, then there is no common interest strong enough to create a CP, and the problem will be instead one of bargaining or arbitration, the outcome of which depends on threat advantage or considerations of fairness. Neither of these provides a conventionalist justification of authority. If, on the other hand, conflict of interest is less than the interest in conformity, the problem remains one of coordination, but the above arguments against the necessity of authority hold.[179]

Green's argument unfolds by: a) assuming at the outset that the best understanding of coordination problems available to conventionalists is the Lewisian one[180], b) stressing that at the core of the Lewisian understanding is the claim that a great deal of similarity between the interests of the agents involved in a coordination problem exists[181], and c) asking us to conclude that, since such a degree of similarity of interests exists, any one of the characteristics that could make an alternative become salient would

[178] Leslie Green, *The Authority of the State*, p. 114.
[179] Leslie Green, *The Authority of the State*, p. 115.
[180] "The logical structure of conventionalist theories can, I think, best be understood by relying on the theory of co-ordination games, as developed by Thomas Schelling, David Lewis and others." (Leslie Green, *The Authority of the State*, p. 95)
[181] While elaborating on Lewis's example about how to settle on a driving convention, Green writes: "Although conventionalism can accommodate conflicts of interests, there must still be a predominance of coincident interests…" (Leslie Green, *The Authority of the State*, p. 95)

be equally worthy and, as such, sufficient to bring about the mutual beliefs, attitudes and expectations that are necessary for the resolution of a coordination problem.

One of the possible responses to Green's argument is to point out that Lewis's understanding of coordination situations is far from being the only one plausible from the point of view of a conventionalist. To begin with, one way of understanding what is entailed by holding that conventions constitute solutions to coordination problems is to say that such an understanding allows us to explain why the fact that other agents are likely to φ provides one with a reason to φ. Conventions are practices that are observed primarily due to there being a mutual expectation (and a corresponding belief) that they are observed. However, by discussing conventions in terms of coordination games one is also shown why (and what it means that) following conventions is the rational thing to do, and not merely a proof of irrational conformity.

Now, let me also suggest that, even if one would consider the Lewisian understanding of coordination problems to be the preeminent one, an appeal to authority might still need to be included in any plausible attempt to solve a coordination problem. Consider a pure coordination problem, in which the only motivation of the parties is that of settling upon a common solution, i.e. upon a common course of action among several available alternatives. Since the parties must, as Green says, "suggest, advise, persuade or threaten" the others in order to reach a solution, the question is which one of them should accomplish this task? If more than one of the parties act towards this end simultaneously, then they become involved in a new coordination problem, having to decide which of the alternative courses of action available will be considered suitable, will achieve the necessary salience and, finally, will become a solution. If this is so, it becomes plausible

to argue, as Eerik Lagerspetz suggests, that an appeal to authoritative practices for finding a solution to a coordination problem presents the important evolutionary advantage of being able to limit the range of relevant sources of information.[182]

However, this view might seem just initially plausible, and easily objectionable. An immediate objection concerns the imminence of an infinite regress of coordination problems, brought about by the need to decide who should be the authority. This is a problem that would arise whenever one has to consider, on the one hand, the pronouncements of the authority (say, the decisions of a court) and, on the other hand, the pronouncements of an alternative source of alleged authority (say, the pronouncements of a gathering of communal elders).

I believe that a comprehensive answer to this objection would have to include a discussion of certain aspects of legal theory that are beyond the scope of this work. While not attempting such an answer, I will sketch a tentative one that I find plausible, by focusing on those aspects which pertain to the present discussion. Let me just point out that this objection seems similar in structure to arguments about the general way in which considerations of salience figure in the discussion of coordination problems and conventionalism. In this sense, one should reply to the objector that a) the reason why the decision of a court is salient in any given case is because the decisions of the courts do enjoy salience in complex contemporary societies, that b) the decisions of courts are salient in complex contemporary societies because they are provided with authority by

[182] Lagerspetz, *The Opposite Mirrors: An Essay on the Conventionalist Theory of Institutions*, p. 95. As a practical example of an authority's limiting the range of legitimate sources of information, Lagerspetz asks us to notice that "in order to find out the content of the existing rules of traffic, I have to consult law-books, or the decision of courts, but not the opinion of all the people I meet on the streets. While doing so, I can rely on the belief that the others have consulted the same sources and expect that I have done the same." (Lagerspetz, *ibid*).

the complex legal systems that exist in complex contemporary societies, that c) these complex legal systems are often (and plausibly) seen as being, at least in part, as the salient solution to the coordination problem that interaction in complex contemporary societies creates. Finally, one should also point out that d) such a salient solution is likely to appear simply because individuals in complex contemporary societies share the practice of adopting salient solutions, a practice which, in turn, is likely rooted in and spread because of its higher efficiency as compared to the practice of not adopting salient solutions at all.

5.6 Beyond Lewis's understanding of convention

Those who believe that moral norms can be analyzed along conventionalist lines and those who attempt to provide a conventionalist account of authority and political obligation share the view that Lewis's definition of convention is too restrictive.[183] Recall that, on Lewis' definition, all agents have approximately the same preferences regarding all possible combinations of outcomes. It should come as no surprise then that those who wish to extend Lewis' analysis of conventions to other contexts – especially moral ones – find his definition overly restrictive. This is because moral norms are invoked in situations in which agents differ in their preferences for combinations of outcomes. According to a prevalent view, the point of moral norms is precisely to regulate situations

[183] See, for instance, Verbeek, *Conventions and Moral Norms: The Legacy of Lewis*, p. 77, Lagerspetz *The Opposite Mirrors: An Essay on the Conventionalist Theory of Institutions*, pp. 45–47, and den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 17.

of potential conflict of interests.[184] We have already seen that Lewis' view allows for imperfect coincidence of interests and that, according to him, coordination problems should be understood as situations of interdependent decision in which coincidence of interests predominates. It should be clear, however, that cases in which the interests of the agents do not overlap fall outside of the scope of Lewis' analysis of convention. Situations that have the structure of a prisoner's dilemma are a case in point.

Dropping the condition which requires that agents have the same preferences regarding all combinations of outcomes may seem, at first glance, a questionable move. Surely, what makes it difficult to apply Lewis' analysis of convention to moral contexts is precisely the above-mentioned condition, so one can object that simply assuming that this is not a necessary condition in order for a convention to exist does not make conventionalist accounts of moral norms plausible. However, as it will soon become clear, dropping the above-mentioned condition is not a dubious move. Let me detail.

The intuition at the core of the attempts to provide a conventionalist account of norms is that the regularities in behavior of the sort analyzed by Lewis could also exist in situations of significant conflict of interest. Hence, the idea that Lewis' analysis of convention can be extended to other contexts. It is important to note, however, that those who argue that the Lewisian treatment of ordinary coordination problems can be extended to other game theoretic situations in a manner which is both plausible and theoretically fruitful, adopt a definition of convention which is different from the one employed by Lewis. Den Hartogh, for instance, states that the basic connotation in his "technical use of the term 'convention' is the existence of interdependent reasons,

[184] See, for instance, Edna Ullmann-Margalit, *The Emergence of Norms*.

reasons derived from a transparent pattern of mutual expectations".[185] Thus, the claim that it is possible to extend the conventionalist analysis of ordinary coordination problems to other game theoretic situations, such as the prisoner's dilemma, amounts to the claim that, even in such situations, rational players can have reason "to make their own choice fit into a pattern of interdependent decisions".[186] Three things need to be stressed here. First, those who attempt to provide a conventionalist account of norms use the term "convention" in a technical sense, which differs both from the ordinary sense of the term and from the technical sense ascribed to it by Lewis. Therefore, arguments to the effect that applying a conventionalist analysis to situations which are characterized by conflict of interests is objectionable miss their point to the extent that they rely on an equivocation between different meanings of term "convention". Second, insofar as conventionalists are right in claiming that basic aspects of the Lewisian conventionalist model are also present in situations characterized by conflict of interests, there seems to be no ground for complaining that stretching the meaning of the term "convention" is inadequate or that dropping the condition which requires similarity in preference is a theoretically dubious move. Third, if conventionalists succeed in establishing that basic aspects of the Lewisian conventionalist model are also present in situations of significant conflict of interests, this is no trivial result. For instance, in the prisoner's dilemma agents seem to have reason to choose their equilibrium strategy regardless of how others choose. Thus, if conventionalists succeed in showing that even in situations that have the structure of a prisoner's dilemma, agents can have interdependent reasons for action, this result is not negligible.

[185] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 17. Compare Verbeek, *The Authority of Norms*, p. 253.
[186] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 6.

So far, I have focused on the idea that Lewis' definition of convention is overly restrictive because it makes similarity in preference a necessary condition for the existence of a convention.[187] Those who attempt to apply Lewis' analysis to moral contexts drop this condition. However, Lewis' definition of convention needs to be further modified if a conventionalist analysis of norms is to succeed. As we have already seen, according to Lewis, a convention is a regularity in the behavior of the members of a given population. Yet, authors like den Hartogh and Verbeek propose to understand conventions as convergent patterns of expectations, rather than convergent patterns of behavior. As Verbeek emphasizes, convergence in behavior can often be taken as an indication that agents are following a norm. Still, a pattern of convergent behavior is not in itself a norm. Moreover, it seems that there can be many reasons why the agents display a converging behavior. These considerations suggest that we should rather think of conventions in terms of the agents' reasons that explain convergence in behavior.[188] Verbeek righty points out that this modification of Lewis' definition of convention is not against the spirit of his analysis. The fact that Lewis places such an emphasis on the idea of common knowledge stands witness to the fact that he intended to offer an account of the agents' reasons for conforming to the convention.[189] The definition of convention in terms of expectations about behavior rather than actual behavior will prove to be important when I will deal with the attempt to provide a conventionalist account of

[187] Lagerspetz, for instance, stresses that Lewis' definition of convention is restrictive in a further sense. On his definition, for a practice to count as conventional the existence of a conventional alternative R' is necessary. However, Lagerspetz argues that a practice may have a conventional character, i.e. its being followed would depend on a mutual belief that others will follow it, even though the alternative for the practice in question would be to have no practice regulating the interactions between agents. See Lagerspetz, *The Opposite Mirrors: An Essay on the Conventionalist Theory of Institutions*, p. 47.

[188] Verbeek, *Conventions and Moral Norms: The Legacy of Lewis*, p. 77, and *The Authority of Norms*, p. 249. See also den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 35.

[189] For a reconstruction of Lewis' argument which supports this interpretation, see Cubitt and Sugden, *Common Knowledge, Salience and Convention: A Reconstruction of David Lewis' Game Theory*.

authority and political obligation. On this account, a convention is a pattern of interdependent expectations that supports convergent behavior.

5.7 Extending the conventionalist analysis from ordinary coordination problems to other game theoretic situations

As I have already pointed out, conventionalists argue that, common assumptions to the contrary notwithstanding, the Lewisian conventionalist analysis can be extended from ordinary coordination problems to other game theoretic situations, such as the prisoner's dilemma. Given that a conventionalist analysis of authority and political obligation can only succeed if it is applicable to situations in which there is considerable conflict of interests, this is a crucial step in their argument. We have already seen that those who attempt to provide conventionalist accounts of norms use a refined version of Lewis' definition of convention. What they claim is that certain basic aspects of the Lewisian conventionalist model are also present in situations of significant conflict of interests. The basic idea is that, even in such situations, a conventional norm could exist such that everyone has reason to conform to it provided that everyone else conforms to it. To claim that such conventional norms could exist in situations that have the structure of a prisoner's dilemma amounts to the claim that it is rational to cooperate in the prisoner's dilemma.

Yet, attempts to establish that it is rational to cooperate in the prisoner's dilemma are typically met with suspicion by game theorists, who stress that on the standard conception of rationality it is straightforwardly rational to defect. The major complaint is

that, given the axioms of revealed preference theory and the insistence of the game theorist that, for any particular game, all utility that is relevant must be built into the game, there is no way of avoiding the conclusion that rational agents will defect in the prisoner's dilemma.[190] Many game theorists consequently point out, just like Ken Binmore, that "it is a tautology that a rational player will defect in a one-shot Prisoner's Dilemma"[191].

In the most popular incarnation of the Prisoner's Dilemma game, we are asked to imagine two persons arrested by the police for committing a crime. While able to convict them both on a minor charge, punishable with a lesser penalty, the police also suspect that the prisoners are guilty of the more serious crime, but they cannot prove it without getting either one of them to confess. Consequently, the police offer each of the prisoners the following deal: confess against your partner in crime and we will let you go free, while demanding a maximum penalty for him; if your partner confesses too, both of you will get punished, with a slightly lesser penalty than the maximum one; if, however, you keep silent and he confesses, it is you who will get the maximum penalty, while your partner goes free; and if neither of you confesses, you will both be convicted for the lesser crime, and get the corresponding lesser penalty.[192] Assuming only that the

[190] Note, however, that those who attempt to establish that it may be rational to cooperate in the prisoner's dilemma by showing that the Lewisian analysis of conventions can be extended from ordinary coordination games to the prisoner's dilemma explicitly reject the assumptions of revealed preference theory. For arguing that the revealed preference theory is theoretically impoverished, see den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, pp. 9, 39, and compare Sugden, *The Motivating Power of Expectations*, pp. 125–126. See also Verbeek, *Instrumental Rationality and Moral Philosophy: An Essay on the Virtues of Cooperation*, chapter 7. For the observation that the traditional game theorist will "insist that all the utility that is relevant to the game must be built into the game", see Gaus, *Reasonable Utility Functions and Playing the Cooperative Way*, section 3.2.

[191] Binmore, *Review of Martin Hollis' "Trust within Reason"*, p. 212. See also, Binmore *Game Theory and the Social Contract, Volume 1: Playing Fair*. For more on this point, see Blackburn, *Ruling Passions*, chapter 6, and Pettit, *The Prisoner's Dilemma Is an Unexploitable Newcomb Problem*.

[192] This is a hugely popular way of depicting the prisoner's dilemma game. In his book *Prisoner's Dilemma*, William Poundstone points out that, while the game as such was discovered by Merrill Flood and

prisoners want to avoid going to jail, each of the will reason as follows. If my partner confesses, I either keep silent and get the maximum penalty, or confess as well and get the slightly reduced penalty. Hence, I better confess too. If, on the other hand, my partner keeps silent, I either keep silent as well and get the lesser penalty for the lesser crime or confess and go free. In this case too, I do better if I confess. Hence, I do better by confessing no matter what my partner does. Since both prisoners reason in the same way, they both confess and end up getting the penalty which is only slightly less than the maximum one.

At the heart of the game are two aspects which are jointly responsible for its dilemmatic character. The first one is that the only equilibrium is one in which each player has one strategy which is strictly dominant (i.e. it dominates all others): for each prisoner, *defection* (which corresponds to confessing) dominates *cooperation* (which corresponds to keeping silent). The second aspect is that the only equilibrium outcome (confess/confess) is Pareto-inefficient – it is, in fact, strongly Pareto-inferior to the non-equilibrium outcome (keep silent/keep silent). While both players would be made strictly better off by playing *cooperation* (i.e. by keeping silent), it is individually rational for each of them to play *defection* (i.e. to confess). This is so because each prisoner has to assume that the other is rational and will consequently confess (this is so because confessing, as we saw, is the single best strategy for each prisoner regardless of what strategy the other chooses) and, on this assumption, his only rational choice is to confess (in order to minimize his own damage).

Melvin Drescher, its name was couched by Albert W. Tucker, "because of a story that he used in order to illustrate the game" (p. 8). My own rendering of the game and certain aspects of the ensuing discussion follow Itzhak Gilboa *Rational Choice* (pp. 91–103), Gerald Gaus *Reasonable Utility Functions and Playing the Cooperative Way* (esp. pp. 218–230), and Bryan R. Routledge *Economics of the Prisoner's Dilemma: A Background* (pp. 92–111).

The idea, then, is that sometimes even rational agents reasoning purely instrumental will end up in PD situations. Therefore, as Robert Aumann puts it:

> *People who fail to cooperate for their own mutual benefit are not necessarily foolish or irrational*; they may be acting perfectly rationally. The sooner we accept this, the sooner we can take steps to design the terms of social intercourse so as to encourage cooperation.[193]

Several interesting questions arise. First, can we legitimately claim that the players should reason differently, so that they avoid ending up locked in a Pareto-inefficient outcome? If so, should we claim that what we are pointing out is a solution to the game, or rather an elaboration on it and, also, can it be the case that, even if there are no resolutions to the normal-form game, attempts to elaborate starting from the prisoner's dilemma can be expected to lead to non-trivial results? If there are non-trivial results to be obtained, will they necessarily point towards a "moral solution" to the dilemma, or should we be rather looking for non-moral elaborations of the game?

In what follows, I will look at an attempt to show how a modified version of the analysis put forward by Lewis can be used to indicate the way in which, even within the setting of the prisoner's dilemma, conventions can emerge which, because they are focused on the Pareto-optimal outcome, will allow the players to avoid ending up locked in a Pareto-inefficient outcome. The attempt I have in mind begins as a series of remarks about Hume's famous example of the two farmers (say, A and B) who must decide whether to help each other reap their harvests.[194] It qualifies Hume's example by making

[193] Aumann, *Game Theory*, p. 468 (emphasis in original).

[194] In the form considered here, this attempt is put forward by Robert Sugden and Bruno Verbeek, but its roots can be partly found in Robert Axelrod's work on the tit-for-tat strategy. See Sugden, *The Economics of Rights, Co-operation and Welfare* (esp. pp. 111–125), Verbeek, *Conventions and Moral Norms: The Legacy of Lewis* (esp. pp. 73-86), Axelrod, *The Emergence of Cooperation Among Egoists* (pp. 306-318), and Axelrod, *The Evolution of Cooperation* (esp. chapter 1). For Hume's example of the two farmers, see Hume, *A Treatise of Human Nature* (III, 2, v).

four plausible assumptions: a) that A and B are in a recurrent situation (in other words, that they are engaged in a repeated prisoner's dilemma); b) that it is at least possible that A and B will meet each other in the same situation; c) that each member of the population to which the two farmers belong will, at least in the long run, gain something by cooperating (hence, that the number of rounds in which the game is played is large enough and that the probability *p* that the two farmers will meet each other again in a similar scenario is $p > 1/3$); finally, d) that the players remember the choices of the other players with whom they have already interacted.[195]

Once these assumptions are made, various cooperative strategies could emerge, among them the Tit-For-Tat (TFT), a strategy of reciprocity centered on the requirement that you should cooperate with those who cooperate with you. Players engaging in TFT begin by cooperating in the first round of the game and replicate each other's choice in every one of the subsequent rounds. Following Axelrod, it could be easily proven that, in a repeated finite prisoner's dilemma, TFT is an equilibrium (i.e. if TFT is followed within the considered population, neither A nor B can improve considering the fact that the other chooses TFT).[196]

Since it is beyond my present purpose to offer a detailed analysis of TFT, let me just mention that, although there are no reasons to prefer to move from TFT to another strategy (since the others are playing TFT and TFT is just as successful as any of the other available strategies), this is not enough to show that TFT could be a Lewisian

[195] Verbeek, *Conventions and Moral Norms: The Legacy of Lewis*, p. 78. See also Axelrod, *The Emergence of Cooperation Among Egoists*, pp. 308–309.

[196] The proof centers around showing how, assuming that B plays TFT and A knows that B does so, it can be proven that TFT is, in fact, the best reply against itself and, consequently, that TFT is an equilibrium. See Axelrod, *The Emergence of Cooperation Among Egoists*, and Axelrod and Hamilton, *The Evolution of Cooperation*. The proof is restated in Sugden, *The Economics of Rights, Co-operation and Welfare*, pp. 114ff, and Verbeek, *Conventions and Moral Norms: The Legacy of Lewis*, p. 79.

convention within the considered population. In fact, TFT does not satisfy condition (4) of Lewis's definition, which requires that almost everyone prefers that any one more conform to (the regularity in behaviour) *R*, on the condition that almost everyone conform to *R*.[197] As Verbeek notices, Lewis's definition "requires that a convention is a uniquely optimal reply against itself"; this is so because:

> Given the assumptions it cannot be the case that each prefers that any one more conforms to TFT when almost everyone conforms to TFT. In a population of TFT players, a strategy of unconditional cooperation C ('always cooperate, no matter what the other has done in the previous round') does as well as TFT and there is no reason for TFT players to prefer others to follow TFT rather than C. In game theoretic terms, TFT is not stable.[198]

Note, moreover, that it is also possible for TFT not to be an equilibrium anymore. As Sugden argues, it is plausible to assume that players will occasionally make mistakes and, whenever this happens, TFT is no longer the best reply to itself.[199] Say A and B play TFT and, while meaning to cooperate, in round *i* A defects by mistake, while B cooperates. In round *i*+1 B will react by defecting, while A will cooperate, since A reacts to B's cooperation from round *i*. In round *i*+2 the choices will be reversed, and so on. Consequently, by strictly following TFT, A and B will become stuck in a sequence of cooperation-defection.[200]

Should this be the case, the better option would be to "snap out" of the sequence of cooperation-defection by reintroducing cooperation in round *i*+2. This, however,

[197] For Lewis's definition of convention, see section 5.4.2.
[198] Verbeek, *Conventions and Moral Norms: The Legacy of Lewis*, p. 79, and notes 29 and 30 (same page).
[199] See Sugden, *The Economics of Rights, Co-operation and Welfare*, pp. 116ff, and Verbeek, *Conventions and Moral Norms: The Legacy of Lewis*, p. 79.
[200] Verbeek also points out that, the next time a mistake is made, A and B would end up in an "always defect" sequence. See Verbeek, *Conventions and Moral Norms: The Legacy of Lewis*, p. 79.

implies considering a version of TFT which, following Sugden, I will call **T1**.[201] **T1** is a strategy which accounts for the possibility that players will occasionally make mistakes by appealing to the concept of being in good standing. **T1** makes it a requirement that a player who is in good standing is entitled to the cooperation of his opponent. As Sugden explains, playing **T1** goes like this:

> At the start of the game both players are treated as being in good standing. A player remains in good standing provided that he always cooperates when **T1** prescribes that he should. If in any round a player defects when **T1** prescribes that he should cooperate, he loses his good standing; he regains his good standing after he has cooperated in one subsequent round. (This is why I call this strategy **T1**; if it took two rounds of cooperation to regain good standing, the strategy would be **T2** and so on.) Given all this, **T1** can be formulated as follows: 'Cooperate if your opponent is in good standing, or if you are not. Otherwise, defect.'[202]

For a player who never mistakes, TFT and **T1** are equivalent to one another. The two strategies only differ in the choices they recommend once the player has defected by mistake. However **T1** is also a stable strategy, while TFT is not. We can show this by looking at the only three possible scenarios that may appear in a given round *i*, namely: 1) Both A and B are in good standing or neither is. Then A will cooperate in *i* and will repeat B's choice in subsequent rounds (TFT); 2) A is in good standing, B is not. A will defect in *i*, and repeat B's choice in subsequent rounds; 3) B is in good standing, A is not. A will cooperate in *i*, and cooperate in *i*+1, and repeat B's choice in subsequent rounds.

Scenario 1) will occur when $i = 1$, i.e. in round 1, when the players should play TFT. Two options are available in Scenario 2). If A cooperates in *i*, in *i*+1, he will end up in Scenario 1). If A defects, in *i*+1 he will still find himself in Scenario 2). This implies that the same choice is the best in *i*, *i*+1, *i*+2 …. Consequently, the only possible

[201] See Sugden, *The Economics of Rights, Co-operation and Welfare*, p. 116.
[202] Sugden, *The Economics of Rights, Co-operation and Welfare*, p. 116.

sequences are: cooperate, cooperate, cooperate … or defect, defect, defect …. Since as implied by the assumption c) above (concerning the incentives attached to mutual cooperation as expressed by $p > 1/3$), the sequence cooperate, cooperate, cooperate … gives a higher utility, A's best choice in Scenario 2) is just as in Scenario 1) to cooperate. In Scenario 3) A is free to defect in I, since in $i+1$ A will end again in Scenario 1) in which the best choice is to cooperate. It follows that **T1** is a stable equilibrium strategy, and also, since it is the unique best reply against itself, it meets condition (4) of Lewis's definition (which TFT was unable to meet). Note, however, that as both Sugden and Verbeek point out, **T1** is not the only possible stable equilibrium allowed by this model. Another such equilibrium is unconditional defection: once A knows that B will keep on defecting regardless of his standing, A's best reply will also be continuous defection. Sugden further qualifies **T1** by writing that: "[…] **T1** is a convention. […] To begin with it is clearly a convention of reciprocity: a person following **T1** is willing to cooperate proving his opponent is willing to do the same".[203]

We can then conclude that **T1** is a convention. The important thing to notice is that **T1** satisfies the requirements specified in Lewis' definition of convention. Since **T1** is adopted in the considered population, the requirement (1) that almost everyone conforms to **T1** is satisfied. Furthermore, given that **T1** is an equilibrium, the requirement (1) that almost everyone expects almost everyone to conform to **T1** is also met. The requirement (3) that requires similarity in preference concerning all possible combination of outcomes is not met, since this requirement has been dropped from the outset. Given that **T1** is an equilibrium for almost all members of the considered population, it follows that the requirement (4) that almost everyone prefers that any one more conform to **T1,**

[203] Sugden, *The Economics of Rights, Co-operation and Welfare*, p. 118.

on the condition that almost everyone conform to **T1** is also satisfied. Finally, since there are other possible stable equilibria besides **T1**, it follows that the requirement (5) that almost everyone would prefer that any one more conform to **Tn**, on the condition that almost everyone conform to **Tn** is met as well. The upshot is that the Lewisian analysis of convention can appropriately be extended to situations characterized by significant conflict of interests.[204] In other words, Lewisian conventions can emerge and stably be sustained in cases which have the structure of a prisoner's dilemma.

[204] For other attempts to account for the rationality of cooperating in the prisoner's dilemma by showing that Lewisian conventions could exist in cases that have the structure of a prisoner's dilemma, see also Sugden's *The Motivating Power of Expectations*, and Brian Skyrms's *Evolution of the Social Contract* and *The Stag Hunt and the Evolution of Social Structure*. Note that, for instance, the conception of normative expectations put forward by Sugden in his *The Motivating Power of Expectations* is different from the one advanced by conventionalists such as den Hartogh or Verbeek, who claim that in order to account for normative expectations one needs to invoke the concept of "cooperative dispositions". As den Hartogh puts it, cooperative dispositions and patterns of (justified) mutual expectations have an internal reference to each other. See his *Mutual Expectations: A Conventionalist Theory of Law*, p. 20, 36. See also Verbeek, *Instrumental Rationality and Moral Philosophy: An Essay on the Virtues of Cooperation*. By contrast, Sugden argues that a plausible account of normative expectations need not rely on any moral presuppositions, but only on assumptions about the human propensity to feel resentment under certain circumstances. He defines resentment as "a sensation or sentiment which compounds disappointment at the frustration of one's expectations with anger and hostility directed at the person who is frustrating (or has frustrated) them" (Sugden, *The Motivating Power of Expectations*, p. 113). Sugden holds that feeling resentment against those who frustrate one's reasonable expectations and feeling aversion towards performing actions that are likely to arouse resentment is a central feature of human psychology. According to his Resentment Hypothesis, aversion to being the focus of resentment can motivate individuals to meet others' expectations about them. Sugden's theory about normative expectations (i.e. expectations that individuals have about other individuals' actions and that have the power to motivate individuals to act in accordance with them) is intended as a unified theory about how the game theoretic apparatus can be adapted to explain both instances of converging behavior that are contrary to self-interest and instance that are not. More specifically, Sugden claims that, while A's beliefs about B's beliefs about A's choice of strategy has to find a place in the conceptual structure of game theory, conventional game theory cannot make room for such higher-order beliefs, given its assumption that players can be directly motivated only by utilities and beliefs about which strategies will be chosen by the other players. Thus, he points out that the objection according to which his attempt to account for the rationality of cooperating in the prisoner's dilemma (by arguing that normative expectations can motivate individuals to cooperate) is unsuccessful because the game he is analyzing is not a *real* prisoner's dilemma, misses its point. Given that Sugden's modification of the assumptions of standard game theory is intended to cover both cases in which players are self-interested and cases in which they are motivated to act in ways that are contrary to self-interest, his theory about normative expectations seems more attractive than the one proposed by those who appeal to the idea of cooperative dispositions. It should be noted, however, that Sugden's theory is not free of difficulties. For instance, Verbeek argues that Sugden cannot do away with the concept of cooperative dispositions given that resentment already presupposes their existences. See Verbeek, *Instrumental Rationality and Moral Philosophy: An Essay on the Virtues of Cooperation*, chapters 2 and 4. In this chapter, I do not deal with the question of whether Sugden is right in holding that a theory of normative expectations need be grounded on any moral assumptions. My aim is to show that, even if we accept the

As I have pointed out, in the case of a normal-form prisoner's dilemma game the dominant rational strategy is defection. This can also be understood as an indication that the player should not conform to the expectations of the other players. Note, however, that for a pattern of mutual expectations to exist at all within a prisoner's dilemma, the expectations involved should be "expectations to behave in a cooperatively virtuous way"[205].

Once the idea of cooperative dispositions (or cooperative virtues) comes into play, the question of whether attempts to account for the rationality of cooperation in the prisoner's dilemma by assuming the existence of such dispositions qualify as "moral solutions" becomes pressing. Two points need to be emphasized here. First, the conventionalist can insist that his treatment of the prisoner's dilemma does not qualify as a moral solution to the game. Alterations of the game can be divided into two categories: changes in the structure of the preferences of the players and changes in the strategy space. However, as Verbeek points out, if we focus on the nature of the reasons for choosing the stable equilibrium strategy, it will turn out that the conventionalist treatment of the game does not fall into any of the above-mentioned categories.[206] According to Verbeek, the reason why the agents opt for a stable equilibrium strategy like **T1** is not that they have moral preferences or that there is moral strategy in their space. Verbeek proposes to account for their choice in terms of interdependent reasons for action.[207]

view that cooperative dispositions are needed in order to explain how certain conventions emerge and reproduce themselves, this does not render the conventionalist position objectionable.

[205] On this point, see Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 28.

[206] Verbeek, *Conventions and Moral Norms: The Legacy of Lewis*, p. 81, n. 37.

[207] For a more nuanced view, see den Hartogh *Mutual Expectations: A Conventionalist Theory of Law*, p.19.

Second, one can argue that assuming that individuals are cooperatively disposed is not an unreasonable assumption to make. Conventionalists who focus on analyzing situations that have the structure of a prisoner's dilemma can draw attention to the fact that the prisoner's dilemma is different from a zero-sum game. Given that the cooperative strategy offers higher payoffs, a motive to cooperate exists. Therefore, it is not unreasonable to claim, as Govert den Hartogh does, that the main reason why players defect in one-shot prisoner's dilemmas "is not the hope for exploitative gains, but the fear of being duped."[208] Once we adopt this standpoint, it becomes less clear why the fact that, on the conventionalist view, cooperatively disposed individuals are playing an assurance game instead of a prisoner's dilemma should count as an objection against conventionalism. Surely, one could argue that conventionalism is objectionable because it claims to solve the prisoner's dilemma, whereas in fact it changes the game into an assurance game, or because it claims to prove that it may be rational to cooperate in the prisoner's dilemma, while this conclusion cannot be arrived at unless it is assumed that individuals have a disposition to behave cooperatively. However, we should keep in mind that assumptions about cooperative dispositions are not smuggled-in assumptions, and that the form of the argument is different, i.e. the conventionalist claims a) that assumptions about cooperative dispositions are not *ad hoc* assumptions, b) that cooperative dispositions make it possible for cooperative strategies to emerge and stably be sustained, and c) that the standard conception of rationality which prompts in a different direction might be flawed. On the conventionalist picture advanced by Verbeek and den Hartogh, cooperative dispositions are invoked to explain how conventions

[208] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 29.

emerge and how their stability is enhanced. However, the motivation to participate in conventional practices is already present.

5.8 Mutual expectations and political obligations

In what follows, I will offer a brief sketch of how a Lewisian inspired conventionalism can deal with the problem of political obligation. While the following considerations are intended to provide the backbone of a conventionalist account of political obligation, the characteristic features of this account will become clearer in subsequent sections as the argument unfolds.

Let me start by stressing that a conventionalist account of authority and political obligation can succeed only if a conventionalist analysis can properly be applied in situations of significant conflict of interests. Therefore, proving that, common assumptions to contrary notwithstanding, the Lewisian analysis can be extended from ordinary coordination problems to other game theoretic situations, such as the prisoner's dilemma, is a crucial step in the argument. Conventionalists claim that certain basic aspects of the Lewisian conventionalist model can also be present in situations of considerable conflict of interests. More specifically, they argue that patterns of convergent behavior and patterns of interdependent expectations that support them can exist in situations in which the Lewisian condition requiring similarity in preference is not satisfied.

The distinguishing feature of conventionalist accounts of norms is that, on such accounts, the reasons for following a conventional norm are interdependent ones, i.e. they

derive from the fact that such norms are generally being followed. Thus, conventionalism can be viewed as an attempt to substantiate the intuition that collectively stable cooperative strategies manage to emerge and maintain themselves precisely because they serve individuals' mutual interests. However, conventionalism attempts to account to not only for the idea that deviant conduct is instrumentally irrational, but also for the idea that such conduct is open to moral criticism. On the conventionalist account, the reasons for following a conventional norm are derived from the existence of a pattern of mutual expectations. The suggestion is that failure to conform to a conventional norm may be not only instrumentally irrational, but also morally objectionable, since not honoring others' expectations would mean letting them down.

The question is whether we should conclude that the very existence of a pattern of mutual expectations is sufficient to generate obligations to act in accordance with others' expectations. One can point out that, as long as we accept the conventionalist suggestion that, in this context, not letting down those who have come to rely on us is the most relevant moral consideration, it is hard to avoid the above-mentioned conclusion. However, given our strong intuition to the effect that other moral considerations are apt to have more weight in determining whether one is under an obligation to honor others' expectations (e.g. considerations having to do with whether reliance on the agent is legitimate and with whether others would be harmed if the agent failed to conform to their expectations), conventionalists want to block this result. It should be noted that conventionalists acknowledge that claiming that the reasons for following a norm derive from the fact that the norm in question is generally being followed opens the way for

objections.[209] The way out of this problem is to argue that patterns of mutual expectations are not immune to moral criticism and that expectations can only assume an obligating character if they are legitimate.

Lewisian inspired conventionalists like Govert den Hartogh and Bruno Verbeek have developed a sophisticated account of norms which illustrates how conventionalism can accommodate the idea that whether expectations assume an obligating character is answerable to moral standards.[210] Den Hartogh, for instance, points out that "the system of our expectations as a whole is structured by an implicit appeal to underlying values and principles", and therefore, we can establish what we may legitimately expect each other to do.[211] Both den Hartogh and Verbeek make crucial use of the concept of cooperative dispositions. The basic intuition behind their version of conventionalism is that what makes it possible for collectively stable cooperative strategies to emerge and maintain themselves is the fact that individuals are cooperatively disposed. In other words, what makes the existence and stability of certain patterns of mutual expectations possible is not only a pattern of interests, but also a widespread disposition to behave cooperatively. What follows from these considerations is that the patterns of mutual expectations and cooperative dispositions are closely connected. As den Hartogh puts it, patterns of expectations and cooperative dispositions should be viewed as having an internal reference to each other.[212] On the one hand, being cooperatively disposed means being prepared to honor others' justified expectations. On the other hand, the existence of

[209] See, for instance, Govert den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 44.
[210] See den Govert den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, and Bruno Verbeek, *Instrumental Rationality and Moral Philosophy: An Essay on the Virtues of Cooperation*, *The Authority of Norms*, and *Conventions and Moral Norms: The Legacy of Lewis*.
[211] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 44.
[212] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, pp. 20, 36.

such expectations is justified by reference to known dispositions to respond to them. Yet, once the idea of cooperative virtues such as trustworthiness or fairness comes into play, we have a standard against which the legitimacy of expectations can be judged. According to den Hartogh, "the system of our mutual expectations as a whole is structured by its underlying form: you expect me to contribute to the common good because it would be untrustworthy or unfair to betray your trust".[213] However, "it is not really untrustworthy or unfair not to honor your expectations, because, for example, you do not really intend to do your share either, or the burdens of production are unfairly distributed".[214]

On the conventionalist view, political obligation is justified by reference to a pattern of mutual expectations. Thus, if a pattern of mutual expectations has been established which requires contributing to a cooperative scheme, each individual has a reason of fairness to honor others' expectations, provided that these expectations are legitimate in the sense discussed above.

5.9 Salience and interdependent reasons

Much controversy in discussions that focus on Lewis' convention and the possibility of extending the model that he proposes to other contexts revolves around the idea of salience. It is important to be clear about the relevance of the idea of salience in the context of the debate about conventionalism and political obligation. To this end, let

[213] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 44.
[214] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 44.

me briefly clarify the relationship between the idea of salience and the interdependence of reasons. As Lagerspetz rightly points out:

> An agent in a coordination problem does not try to find the alternative which is salient for him; he is searching for the alternative which would be perceived as salient by others. Thus the search for salient alternatives is itself based on mutual beliefs concerning the perception of others.[215]

It should be noted, however, that a lot depends on how exactly we understand the idea of mutual beliefs. Let me detail.

Returning to the framework of an iterated game of pure coordination characteristic of Lewisian conventions, let us consider why the agents expect each other to converge on a particular equilibrium among all the coordination equilibria available. While one cannot exclude the possibility that they could succeed in reaching the equilibrium by sheer luck, i.e. without paying attention to the expected actions of the other agents in deciding how to choose, in the most likely scenario they would succeed by paying attention precisely to that. As Lewis puts it, "they are more likely to succeed – if they do – through the agency of a system of suitably concordant mutual expectations".[216]

According to the reconstruction of Lewis' argument proposed by Govert den Hartogh[217], an agent's reasoning would go like this: when I conclude that you will go to

[215] Lagerspetz, *The Opposite Mirrors: An Essay on the Conventionalist Theory of Institutions*, pp. 47–48.
[216] Lewis, *Convention: A Philosophical Study*, p. 25.
[217] When considering Lewisian conventionalism, the question of how best to account for the emergence and maintenance of conventions and the question of whether the most tenable such account can be attributed to Lewis, should be kept distinct. My aim here is not to establish whether den Hartogh's reconstruction of Lewis' argument is accurate. (Compare, for instance, Robin Cubitt and Robert Sugden, *Common Knowledge, Salience and Convention: A Reconstruction of David Lewis' Game Theory*.) What interests me here is den Hartogh's view that in order for salience to give one a reason to choose an alternative, we need to assume that what is salient is already the object of common knowledge. For a similar view, see also Verbeek, *Conventions and Moral Norms: The Legacy of Lewis*. For the controversy about whether salience should be taken as primitive, see also Verbeek *Instrumental Rationality and Moral Philosophy: An Essay*

A, I reach that conclusion based on something that indicates to me that you will go to A. This "something" will be the fact that A is salient the first time we play, the fact (which becomes a precedent) that we have met at A before the following time, and the fact that we meet with regularity at A on every subsequent occasion once a convention is established. The fact that you will go to A constitutes a reason for me to go to A as well and I also understand that you know a) that this very fact makes me believe that you will go to A and b) that this gives me a reason to go to A as well. In turn, a) and b) provide you with a supplementary reason to go to A as well, apart from the original reason that you had to go to A, which is that "something" indicated to you that I will go to A. In Lewis' words, "in our subsequent reasoning we are windowless monads doing our best to mirror each other, mirror each other mirroring each other and so on".[218] The end result is a system of mutual beliefs about mutual beliefs about mutual beliefs … regarding our both going to A, beliefs that, as den Hartogh puts it, derive from the same "indicating fact".

Yet, there are reasons to think that this view about the structure of the pattern of mutual beliefs in coordination problems is misguided. For instance, den Hartogh claims that Lewis's use of the metaphor of the "windowless monad" stands witness to the fact that he has failed to take seriously precisely the idea of interdependent reasons. As he points out:

> The idea of an independent basis of expectation is really a rudimentary form of parametric reasoning in a strategic context: it presupposes that it is

on the Virtues of Cooperation, xi. Compare Robert Sugden, *The Economics of Rights, Co-operation, and Welfare*, and Robin Cubitt and Robert Sugden, *Common Knowledge, Salience and Convention: A Reconstruction of David Lewis' Game Theory*.

[218] Lewis, *Convention: A Philosophical Study*, p. 38.

> always possible to assign independent probabilities to the choices of each of the players. Of course, it is *possible* that one of them has an independent reason to choose one of his alternatives, but in that case it cannot really be a game of pure coordination they are playing.[219]

Den Hartogh claims that, on Lewis's view, the entire pattern of beliefs of lower and higher order is supposed to derive from the same "indicating fact". To see why this view is misguided, consider the following dilemma.[220] If, on the one hand, the indicating fact is sufficient for me to conclude that you will go to A, then I have a decisive reason to go to A myself. This renders the entire superstructure of higher-order beliefs redundant. What follows is that we are not faced with a problem of coordination after all. If you have an independent reason to go to A, and this is sufficient for me to identify the course of action which is rational for me, then the solution is individually accessible. If, on the other hand, I cannot confidently conclude from the indicating fact that you will go to A, then I am unable to conclude that you know that I have reason to go to A. In other words, the superstructure of higher-order beliefs cannot even get off the ground. The upshot of these considerations is that, if a coordination problem can be said to exist, any "indicating fact" on which choice can be based must indicate transparently, i.e. its indicating force must be common knowledge.[221]

A distinction between independent salience and interdependent salience may help further clarify the issue at stake. Den Hartogh illustrates the idea of independent salience by pointing out that, if I expect you to go to A on account of certain information about your biography, there is no coordination problem, since the solution is individually accessible. Admittedly, it is interdependent salience that is relevant in a coordination

[219] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 33, emphasis in original.
[220] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, pp. 33–34.
[221] For a similar line of argument, see Verbeek, *Conventions and Moral Norms: The Legacy of Lewis*, p 83.

problem. When we are in a coordination problem we are looking for "what we mutually know will be salient for each of us".[222]

Three points need to be emphasized here. First, what follows from accepting that any "indicating fact" must indicate transparently, and therefore, the indicating force of any such fact must already be common knowledge, is that salience itself is a matter of convention.[223] Yet, if this is the case, it seems that conventionalism cannot provide an account of how conventions emerge, but only an account of how they reproduce themselves. I will come back to this point later on.

Second, note that, according to den Hartogh, the implication of the argument considered above is that the idea that a pattern of mutual expectations should be justified on an independent basis must be given up. Den Hartogh argues that:

> We know that other road-users expect us to keep to the right; we have no reason to change this belief as long as they keep to the right themselves. We do not need to know any independent reasons for their expectations. But if we do know their expectations, we know what to decide ourselves. Their expectations are a sufficient ground for our decision. And hence for *their* expectations! Everyone has a reason to do what the other, for whatever reason, expects him to do. And precisely that fact is sufficient ground for everybody to continue in their expectations. Nobody needs more grounds. When a pattern of expectations has been established, on whatever basis, then it is self-perpetuating without our ever having to return to that basis again.[224]

Third, the idea that patterns of mutual expectations are not justified on an independent basis has far-reaching consequences. On den Hartogh's account, the fact that one has a disposition to act fairly is not an independent reason to behave cooperatively in

[222] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 34.
[223] For emphasizing that what is salient is a matter of convention, see also Verbeek, *Instrumental Rationality and Moral Philosophy: An Essay on the Virtues of Cooperation*, p. 51. See also Cubitt and Sugden, *Common Knowledge, Salience and Convention: A Reconstruction of David Lewis' Game Theory*, pp. 202–203, for stressing the need to account for the co-evolution of conventions and conceptions of salience.
[224] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 34, emphasis in original.

situations that have the structure of an assurance game or a prisoner's dilemma. To the extent that fairness prompts one to make the cooperative choice, it does so because one is already expected to behave cooperatively. However, den Hartogh points out that this does not mean that one is expected to behave cooperatively on account of any other fact. The reason for expecting one to act cooperatively is knowing that one will not let others down. As den Hartogh puts it: "The cooperative virtue and the pattern of expectations have an internal reference to each other. There is no independent basis."[225]

Before going any further, let me make a few remarks about whether salience is supposed to play a role in explaining the emergence of conventions or, rather, their maintenance. There seems to be fairly wide agreement that salience plays a key role in explaining the *maintenance* of conventions.[226] There is less agreement, however, about whether Lewis should be interpreted as adopting this stance.

For instance, Robin Cubitt and Robert Sugden argue that, although salience is invoked by Lewis in the context of discussing solutions to one-off coordination problems, he is not primarily concerned with the origin of conventions. Lewis brings into discussion Schelling's experiments in order to prove that prior communication is not necessary for solving one-off coordination problems.[227] The subjects in Schelling's experiments manage to coordinate because they pick the salient equilibrium. It should be noted, however, that, according Lewis, the sort of coordination games that are the object of

[225] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 36.
[226] See, for instance, Lagerspetz, *The Opposite Mirrors: An Essay on the Conventionalist Theory of Institutions*, Verbeek, *Instrumental Rationality and Moral Philosophy: An Essay on the Virtues of Cooperation*, and Cubitt and Sugden, *Common Knowledge, Salience and Convention: A Reconstruction of David Lewis' Game Theory*.
[227] Lewis, *Convention: A Philosophical Study*, pp. 35–36.

Schelling's experiments, i.e. self-contained games, are highly unusual. The reason why Lewis invokes them is that they promise to shed light on how more common coordination problems can be solved without communication. His suggestion is that, confronted with familiar coordination problems, individuals have the tendency to follow *precedent*. And precedent is to be understood as a form of salience. In other words, what helps individuals solve recurrent coordination problems is the salience of precedent. So it seems that salience plays a key role in explaining how conventions tend to maintain themselves.

Some authors disagree with this interpretation of Lewis's view on the role of salience.[228] It is also worth pointing out that whether certain objections against Lewis's view on conventions are damaging will depend on whether we read him as holding that salience plays a part mainly in explaining how conventions reproduce themselves. Consider, for instance, Gilbert's argument to the effect that salience cannot facilitate successful coordination since rational agents have difficulties in recognizing it, and even if they did not, salience would fail to guide them.[229] As some commentators rightly stress, this objection is valid only if we make several assumptions, i.e. the agents are faced with a one-off coordination problem, they do not possess any relevant background knowledge, and there are only two equally good alternatives.[230] In such cases, even if one alternative is perceived as salient, it is not entirely clearly that agents have a reason to choose it. Suppose that one of the available alternatives has a property which makes it salient. One

[228] For instance, the argument put forward by Cubitt and Sugden is directed against the interpretation of Lewis proposed by Brian Skyrms, according to which the role of salience is to solve the problem of selection between alternative conventions. See Brian Skyrms, *Evolution of the Social Contract*.
[229] Gilbert, *Rationality and Salience*.
[230] Lagerspetz, *The Opposite Mirrors: An Essay on the Conventionalist Theory of Institutions*, pp. 48–49, and Miller, *Co-ordination, Salience and Rationality*.

can wonder why is the fact that one alternative has this property more important than the fact that the other alternative lacks it. Why choose the former and not the latter? Yet, if there are more than two alternatives and only one of them is perceived as salient, agents would presumably have a reason to choose that alternative. Moreover, if the agents are faced with a coordination problem with which they have been confronted before, then salience can guide their choices. Admittedly, in such cases the agents have a reason to follow precedent.[231] Thus, if we concede that one-off coordination problems in which the agents possess no relevant background knowledge are uncommon, salience does seem to play an important role in facilitating successful coordination.

If the role of salience is limited to explaining how conventions reproduce themselves, the question is what explains the emergence of conventions. It should be stressed, however, that several leading conventionalists explicitly state that they do not aim to offer an account of the origin of conventions, but simply an account of how

[231] One can insist that, even in cases when what is supposed to facilitate successful coordination is the salience of precedent, the problem identified by Gilbert is still present. For more on this point, see Gilbert, *On Social Facts*, pp. 332–333. To stick to Lewis' example of the interrupted telephone call, Gilbert argues that calling back because this is what the original caller did on the previous occasion is just as unique as not calling back because this is not what the original caller did on the previous occasion. Gilbert concludes that: "Were one faced with a rational Martian, one would hardly know, on the basis of known rationality, what kind of psychological set one was dealing with. We can conclude that in general it is not the case that salience of a precedented combination of actions is generated by common knowledge of successful precedent, given common knowledge of rationality" (*ibid*., p. 333). For a possible answer to Gilbert's objection, see Cubitt and Sugden, *Common Knowledge, Salience and Convention: A Reconstruction of David Lewis' Game Theory*. Cubitt and Sugden maintain that common knowledge in Lewis's sense is possible only if agents have reasons to believe that they posses relevant background information and common standards of inductive inference. According to them, these inductive standards are grounded in common conceptions of what is to count as a "natural analogy". On the interpretation of Lewis proposed by Cubitt and Sugden, Gilbert's point has already been conceded. According to Cubitt and Sugden, a distinctive feature of Lewis' account is that he "treats each instance of a recurrent coordination problem as posing its own problem of equilibrium selection" (*ibid*., p. 201). This is precisely why he is interested in how coordination equilibria reproduce themselves. However, unlike later game theorists, Lewis acknowledges a problematic aspect of real-world coordination problems, i.e. that no two interactions are alike in all respects. Thus, as Cubitt and Sugden point out, "the idea of 'repeating what was done in previous instances of the game' is not well-defined", and therefore, precedent will have depend on analogy (*ibid*., p. 196). On Lewis' account, the solution to this problem is to assume common standards of inductive inference. The idea of salience (i.e. the idea that certain regularities are privileged or "projectible", or that there are "natural analogies") is invoked in order to justify these assumptions.

conventions, once emerged, maintain themselves. Den Hartogh, for instance, argues that the two main elements of his account of conventions, i.e. patterns of mutual expectations and cooperative dispositions, should be understood as having an internal reference to each other. Being cooperatively disposed means being prepared to honor each other's justified expectations, whereas the existence of such expectations is justified by reference to known dispositions to respond to them. Den Hartogh points out that a significant corollary of this fact is that "the mutual expectations of the people participating in a social norm cannot have developed independently of any pre-existing expectations". He goes on to stress that "only if the pattern of expectations already exists in a general way, is it possible to form concrete expectations of behavior in any particular case".[232] What this implies is that his account can only explain the maintenance, as opposed to the emergence, of conventions, a limitation that den Hartogh is ready to accept. Alternatively, one can point out that, since a proper analysis will reveal that what is salient is a matter of convention, the conventionalist has to show how this seemingly vicious regress can be escaped. According to Verbeek, the conventionalist should resist the claim that this is a vicious regress, and argue that, although his theory implies that conventions arise against the background of other conventions, the problem of how conventions emerge in the first place cannot be accounted by it.[233] Finally, Cubitt and Sugden maintain that there is no reason why the fact that conventionalism leaves open the question of the origin of conventions should count as an argument against it, as long as it

[232] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 20.
[233] Bruno Verbeek, *Instrumental Rationality and Moral Philosophy: An Essay on the Virtues of Cooperation*, p. 51. Verbeek points out that this strategy "limits the ambition and scope of conventionalism considerably, but in the end this may be the only tenable form of conventionalism".

succeeds in offering an adequate account of the nature of conventions.[234] All these authors converge on the idea that the question of how conventions emerge is far less important than the question of how to understand their nature and the question of how they can persist over time. [235] On their view, the aim of conventionalism is not provide an account of the origins of norms, but simply to offer an analysis of norms which, to paraphrase Lewis, permits norms to be conventional. (Note further that, according to Verbeek, what prompts the question of how to understand the emergence of norms is a worry about whether the conventionalist can come up with an explanation of their emergence consistent with the explanation of their stability. On the conventionalist view, the stability of norms is explained from the point of view of the individual. Yet, in order for this explanation to be consistent in its methodologically individualist assumptions, it must be the case that norms "could emerge, not necessarily that they actually did emerge, through the interactions between individual agents".[236] Conventionalists attempt to discharge the above-mentioned task by arguing at length that conventions or norms could emerge in cases that satisfy the requirements of methodological individualism.[237] We should keep in mind, however, that conventionalism can only address the question of how a particular convention or norm came into existence.[238])

[234] Cubitt and Sugden, *Common Knowledge, Salience and Convention: A Reconstruction of David Lewis' Game Theory*, p. 203.
[235] Things look different from the perspective of modern evolutionary game theory. If one adopts this perspective, an account of the dynamic processes by which equilibria emerge seems necessary. This might explain why an author like Brian Skyrms, who uses the methods of evolutionary game theory instead of those of classical game theory in modeling the evolution of social norms, tends to conclude that salience is redundant as soon as it turns out that it cannot play a crucial role in explaining the emergence of equilibria.
[236] Verbeek, *Instrumental Rationality and Moral Philosophy: An Essay on the Virtues of Cooperation*, p. 29. Compare den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 7.
[237] Verbeek, *Conventions and Moral Norms: The Legacy of Lewis*.
[238] Verbeek, *Instrumental Rationality and Moral Philosophy: An Essay on the Virtues of Cooperation*, p. 51.

5.10 Objections to conventionalist accounts of authority and political obligation: some replies

We are now in a position to show that conventionalism has the resources to meet the most compelling objections raised against conventionalist accounts of authority and political obligation.

Consider the most common, and apparently the most damaging, objection against such accounts. According to it, a conventionalist analysis cannot properly be applied in situations of considerable conflict of interests. The proponents of this objection point out that a Lewisian convention is arbitrary. What characterizes a Lewisian convention is the existence of more than one possible coordination equilibrium. In other words, Lewisian conventions arise in circumstances when everyone would have equally good reasons to conform to an alternative convention. To stick to a commonly used example, although one may have reason to drive on the left as long as one expects everyone else to do so, one might just as well have reason to drive on the right if one expected everyone else to do so. Yet, critics point out that attempts to provide a conventionalist account of authority and political obligation can only succeed if a conventionalist analysis is applicable in situations of significant conflict of interests.

As I have already pointed out that, the proponent of conventionalist can reply by drawing attention to the fact that those who attempt to provide a conventionalist account of norms typically use the term "convention" in a technical sense, which differs both from the ordinary sense of the term and from the technical sense ascribed to it by Lewis. Consequently, in order to determine whether a conventionalist analysis can properly be

applied in situations of significant conflict of interests, one has to make sure that it is the relevant sense of the term "convention" that guides the inquiry. We have seen that those who believe that norms can be analyzed along conventionalist lines typically use the term "convention" to refer to a pattern of interdependent expectations. Thus, in order to rebut the objection that a conventionalist analysis can properly be applied in situations of significant conflict of interests, it suffices to show that a pattern of convergent behavior and a pattern of interdependent expectations can exist in situations in which the Lewisian condition requiring similarity in preference is not satisfied. This is precisely what conventionalists attempt to do in arguing that the treatment of ordinary coordination problems can be extended to other game theoretic situations, such as the prisoner's dilemma.

There are, however, several considerations that can be added here. One can draw attention to the fact that interdependent reasons for action can only exist if more than one stable equilibrium could have emerged. Thus, on a conventionalist view, although it is the case that a certain pattern of interdependent expectations has emerged, it is also the case that a different pattern of interdependent expectations could have emerged. The conclusion that arbitrariness is an essential ingredient of conventionalism seems unavoidable. It is important to stress that, if it is conceded that any pattern of interdependent expectations could have emerged, this would have troubling consequences for the conventionalist account of political obligation.

However, the proponent of conventionalism can insist that the element of arbitrariness that is built in any conventionalist theory is less damaging to his view than it may at first seem. According to Verbeek, there are several considerations that can be

adduced in favor of conventionalism.[239] First, one can point out that, although it is the case that more than one pattern of interdependent expectations could have emerged, it is certainly not the case that *any* such pattern could have emerged. Given that only some patterns of interdependent expectations can be *stable*, most patterns are excluded. Second, one can stress that some practices do not rely on *interdependent* expectations. Slavery is a case in point. As Verbeek rightly points out:

> It may very well be that the masters expect the slaves to serve them and that the slaves expect that the masters expect them to serve. However, the expectations of masters and slaves do not depend on each other as in the case of property. The slaves have reason to serve the masters because they will be beaten if otherwise. Their reason for serving the master is independent of the expectations of the masters.[240]

Finally, one can draw attention to the fact that, according to conventionalists, the relevant patterns of interdependent expectations cannot emerge unless a sufficient number of agents are cooperatively disposed. Assuming that conventionalists are right in claiming that cooperative dispositions are needed for such patterns to become stable, this would again limit the range of the patterns of interdependent expectations that could in fact emerge. For instance, a pattern according to which relying on others to do their part would be unfair could not rely on cooperative dispositions for its stability.

Now, let me briefly consider Green's argument to the effect that attempts to provide a conventionalist account of authority are conceptually incoherent.[241] According to Green, the conventionalist must view authority as the necessary means to solving

[239] Verbeek, *The Authority of Norms*, p. 253–254.
[240] Verbeek, *The Authority of Norms*, p. 253.
[241] Green, *The Authority of the State*, pp. 113–115, 118.

coordination problems. More specifically, authority is supposed to secure solutions to coordination problems by making one of the available alternatives salient. However, this is problematic given the account of the nature of authority that the conventionalist is presumed to hold. On this account, authority provides exclusionary reasons for action. Yet, Green argues that authoritative decisions cannot function as exclusionary reasons since, in coordination problems, no excluded reasons could exist. This is because authority is needed to solve coordination problems precisely when there seem to be no other salient clues that can be followed.

In responding to Green's challenge, one strategy available to the conventionalist is to raise doubts about whether Raz's conception of exclusionary reasons is the only viable model of the reasons that individuals have for complying with authoritative directives. This strategy is pursued by den Hartogh.[242] My aim is not to offer a conclusive answer to question of whether den Hartogh succeeds in establishing that Raz's conception of exclusionary reasons should be abandoned. Nonetheless, I want to make the more modest suggestion that, insofar as the conventionalist take on the idea of interdependent reasons is found plausible, the outline of a model of the reasons for complying with authoritative directives is within our reach. One way to substantiate the intuition that, if individuals have sufficient reasons to comply with authoritative directives, these reasons can only be accounted for in terms of interdependent reasons for action would be to claim, as McMahon suggests, that individuals have a reason for compliance, a reason that is individually accessible, but accessible only via the principle of collective rationality. On McMahon's view, the principle of collective rationality can give one a reason for compliance that one would not have if one considered the matter merely from the

[242] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, chapter 7.

perspective of the principle of individual rationality. Another way to articulate the above-mentioned intuition is to maintain, as conventionalists suggest, that individuals have an interdependent reason for compliance that is derived from a pattern of mutual expectations. If the concept of exclusionary reasons is intended to account for the idea of being bound to follow authoritative directives, the concept of interdependent reasons seems also well-suited to account for this idea. Admittedly, if the concept of exclusionary reasons is useful in explaining the nature of the reasons to comply with authoritative directives precisely because an exclusionary reason can be distinguished from a mere reason of significant weight, the same holds for the concept of interdependent reasons. To claim that there are interdependent reasons for compliance is to claim that such reasons are justified from a standpoint different from the strictly individualistic one, and in this sense an interdependent reason can be distinguished from a mere reason of significant weight.

A further objection against conventionalist accounts of political obligation is that such accounts fail to capture the categorical nature of obligation.[243] The proponents of this objection point out that the reason to act in conformity with a convention is conditional on the agent's relevant interests. In other words, conventions are hypothetical imperatives, and thus, they cannot engender obligations.

There are several lines of response to this objection. First, one can point out that it would wrong to think that conventional norms have only hypothetical authority, since the existence of a convention can provide one with a reason to follow it, even though

[243] See, for instance, Green, *The Authority of the State*, chapter 4.

following it would be against one's interest.[244] We should not lose sight of the fact that, on the conventionalist account that concerns us here, what makes the existence and stability of conventions possible is not only a pattern of interests, but also a widespread disposition to behave cooperatively. In order for a pattern of interdependent expectations to maintain itself, the agents must be disposed not to let others down, even sometimes this would requires going against one's interest. Thus, a conventionalist analysis that relies on the idea that a widespread cooperative disposition may be needed in order to further individuals' mutual interests can account for the idea that, in following a convention, agents can sometimes act against their interest.

Second, one can draw attention to the fact that conventionalism does not imply that deviating conduct is immune to moral criticism.[245] Once the idea of cooperation disposition comes into play, it becomes apparent how conventionalism can make sense of the idea that deviating conduct can be criticized on grounds other being instrumental irrational. On the picture according to which mutual expectations and cooperative dispositions have an internal reference to each other, failing to honor others' expectations can be criticized based on considerations having to do with the principle of reliability.

Third, one can stress that, although conventionalism offers a naturalistic view of obligation, in the sense that the obligation to act in conformity with a conventional norm is grounded in the existence of a pattern of mutual expectations, which is a contingent social fact, this does not imply that such expectations are not open to moral criticism.[246]

[244] For this line of argument, see Verbeek, *The Authority of Norms*, pp. 250–252.
[245] See, for instance, den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, pp. 17ff.
[246] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 44.

Conventionalists hold that in order for expectations to generate obligations, they must be legitimate.[247] I will come back to this point later on.

Fourth, one can argue that, even though conventionalism cannot account for the unconditional nature of political obligation, this should be viewed as a strength rather than a weakness of the theory. According to den Hartogh, for instance, to deny that political obligation can only be a conditional and selective duty means to misunderstand its nature.[248] However, such denials are common in the literature. By contrast, den Hartogh contends that, since the "existence of obligations to authority does not belong to the rock-bottom of moral fact", the obligation to comply with authoritative directives is always conditional.[249] Yet, one can claim that the view according to which the obligation to comply with the law is unconditional goes hand in hand with the view that the reasons for compliance are best characterized as exclusionary reasons, and that as long as we have not been given any reason to doubt that this is the most adequate characterization, we should stick to the above-mentioned view on the nature of political obligation. While an elaborate treatment of the issue of whether we should abandon the view that the reasons for compliance are to be understood as exclusionary reasons is beyond the scope of this work, I have nevertheless suggested that, if conventionalist accounts of authority and political obligation are incompatible with the aforementioned view, this is less damaging than it may at first seem, since conventionalists can offer an alternative characterization of the reasons for compliance. If it turns out that conventionalism is both plausible and coherent, such an alternative is worth exploring in more detail. A further consideration that can be adduced in defense of the claim that conventionalism does not

[247] On this point, see also Verbeek, *The Authority of Norms*, pp. 251–252.
[248] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, chapter 5.
[249] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 105.

misrepresent the categorical nature of political obligation has to do with the selective nature of this obligation. Den Hartogh argues that, just as the authority of a captain of a life-boat does not extend over actions in a certain category, certain actions cannot properly be commanded by political authority. More concretely, he calls attention to the fact that, arguing that a given principle can engender a political obligation and arguing that political obligation is suitably general are two separate enterprises, and that, for any such principle, we should expect that it does not apply in all circumstances in which political authority demands obedience.

By way of concluding this discussion, I want to add a few remarks about whether conventionalism has the resources to show that one can have a moral reason to obey the law.[250] As I have already pointed out, even staunch anti-conventionalists like Leslie Green admit that conventions constitute an interesting venue for the study of authority and political obligations given that they lack the problematic aspect of enforcement. One of the attractive features of a conventionalist analysis is precisely its instrumentalist character. Conventions are to be viewed as means to promote the mutual interests of individuals. Even though conventionalism allows that such interests are not optimally served (i.e. it allows that some are better off while no one is worse off), conventions can only exist if there is no temptation for individuals to defect. Thus, conventions do not have to be enforced (or, to use the terminology of game theory, conventions are stable). However, we should not lose sight of the fact that those who attempt to account for the obligation to obey the law are looking for moral reasons for obedience. Conventionalism

[250] For an interesting treatment of the question of whether one can have a moral reason to follow a convention, see also Andrei Marmor, *Social Conventions: From Language to Law*, chapter 6.

is typically dismissed as unable to provide such reasons due to its being a normative framework which is instrumentalist in character.

Nevertheless, the question of whether one can have a moral reason to follow a convention is worth considering in more detail. On the conventionalist account that has been the focus of this chapter, one's reason to comply with a convention X is that others expect one to do so. To put it differently, one's reason to comply with X is precisely that X is an established convention. At first glance, this reading seems to rule out the existence of a moral reason for compliance. Yet, if expectations can generate genuine obligations, then we are led to accept the possibility of there being such a moral reason. If den Hartogh's view about how mutual expectations engender obligations is vindicated, to say that an individual has reason to comply with an established convention is to say that others rely on him to do as they expect and that there is reason not to let others down.

Philosophical anarchists concede that one can have a moral reason to obey the law if a) there is an independent moral justification for acting as the law requires, or b) the law helps to secure independently desirable coordination.[251] However, they remain skeptical about the possibility of establishing that one can have a moral reason to obey the law qua law. In what follows, I will consider the question of how a proponent of den Hartogh's version of conventionalism can respond to the above-mentioned challenge. As I have already pointed out, conventionalists explicitly state that existing expectations are not immune to moral criticism.[252] Moreover, it should be emphasized that, when discussing the issue of whether authority provides exclusionary reasons for action, den Hartogh insists that, while authority provides content-independent reasons for action, it

[251] See Simmons, *Moral Principles and Political Obligations*, p. 194.
[252] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, pp. 44, 109–112.

does not require a surrender of judgment.[253] More specifically, he argues that there is nothing in the concept of content-independent reasons that implies that such reasons cannot be weighed against other substantial reasons and ultimately be defeated. If we consider the question of how the conventionalist can respond to the anarchist's challenge, the suggestion that the individual can arrive, on the basis of substantial moral reasons, at the judgment that reliance on him to do as expected is not legitimate, will complicate the picture. Let me explain.

The most straightforward way to show that, on the conventionalist account, one can have an independent moral reason to obey the law is to argue that, when the law is aimed at providing solutions to coordination problems, one can have a moral reason to comply with it, a reason that is different from the moral reason that individuals might have had for seeking a solution to that problem in the first place.[254] The conventionalist can point out that, once a pattern of mutual expectations has been established, one has a moral reason to honor others' expectations. That this is a moral reason becomes apparent as soon as we think about the role of cooperative virtues in prompting one to act in accordance with others' expectations. As den Hartogh puts it, the reason why individuals behave cooperatively is that they feel constrained by fairness or fidelity once certain expectations exist.[255]

Furthermore, it can be argued that the above-mentioned reason counts as an independent moral reason to comply with the law. Assuming that one believes that the reasons that have initially justified the exercise of coordinating authority are no longer

[253] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, chapter 7.
[254] As Marmor rightly emphasizes, individuals can have a moral reason to solve a coordination problem. See Marmor, *Social Conventions: From Language to Law*, p. 134.
[255] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 36.

valid, one may still have a reason to act as required on the basis of the principle of reliability. However, given that conventionalists make room for the idea that expectations are not immune to moral criticism, one can raise doubts about whether the reason to act in accordance with authoritative directives can really count as an independent moral reason. Once we accept the possibility that the individual may judge that certain expectations are not legitimate, and therefore, should not be honored, the following dilemma seems to arise. It is either the case that others' expectations provide one with an independent moral reason to act in conformity with them or it is not. If we choose the first horn of the dilemma, then it cannot be the case that determining whether one should ultimately conform to others' expectations rests on antecedent moral reasons that determine whether these expectations are legitimate. If, however, such antecedent moral reasons play a crucial role in determining whether others' expectations should be honored, we must take the second horn of the dilemma.

The key to this apparent dilemma is to point out that, even though the individual may judge, based on antecedent moral reasons, that certain expectations are not legitimate, and therefore, should not be honored, this does not rule out the possibility of there being an independent moral reason to obey the law. What the conventionalist has to establish is that at least in some cases one has an independent moral reason to act as required by the law. Consider, first, legal prohibitions concerning murder of theft. Surely, conventionalists do not want to claim that the only reason to refrain from violating such prohibitions is that everyone expects everyone else to do so. Verbeek, for instance, stresses that to claim that conventionalism provides a powerful instrument in the analysis of norms does not mean to claim that "everything we can say about our obligations and

reasons for action can be explained by conventionalism".[256] Furthermore, note that den Hartogh explicitly states that mutual expectations constitute a necessary, but not a sufficient condition for the existence of political obligation.[257] According to him, political obligation, if it exists, can only be accounted for in terms of interdependent reasons for action. While this means that whether people commonly believe that they have such obligations has considerable importance, such mutual beliefs are not sufficient for the existence of political obligation. Den Hartogh argues that political obligation is to be understood a non-basic duty. On this view, there is nothing wrong with invoking moral principles in order to determine whether the conditions are under which one can have an obligation to act as required by the law are satisfied.

Now let us focus on situations in which the law is aimed at providing solutions to coordination problems. The conventionalist can insist that at least in some of these situations one can have an independent moral reason to comply with the law (i.e. a reason that is different from the moral reason that individuals might have had for seeking a solution to that problem in the first place). Consider the following excerpt from den Hartogh:

> The circumstances under which coordinating authority is valid may, however, undergo subtle changes, and the preferences of the people involved may shift just as subtly. The moment may come that it is no longer your first priority to bring your course of action into line with that of the others, while your previous conduct had been a ground for them to count on you. At that moment, it is a requirement of *reliability* not to let others down, not to damage their confidence in you.[258]

[256] Verbeek, *The Authority of Norms*, p. 255. Verbeek argues that (a correct characterization of) the norm against killing innocent people would apply authoritatively even in the hypothetical situation in which no one abstained from murder.
[257] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, pp. 109–112.
[258] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 120.

Admittedly, in a situation like the one described above one would have an independent moral reason to comply with law. The conventionalist can emphasize a) that the reason for compliance is generated by the principle of reliability, and therefore, it is a moral reason, and b) that the considerations in favor of compliance are different from the considerations that made the exercise of coordinating authority desirable in the first place, and thus, provide one with an independent reason to obey the law.

5.11 The order of justification within conventionalist accounts of political obligation

In the remainder of this section, I will briefly address two related worries. The first worry has to do with whether the order of justification within a conventionalist account of political obligation is appropriate. The second worry concerns the claim that it is rational to cooperate in prisoner's dilemmas.

Let me start by briefly going back to McMahon's attempt to justify political obligation by reference to the principle of collective rationality (PCR). Recall that McMahon starts from the observation that many people do in fact contribute to cooperative schemes that have the structure of multi-person prisoner's dilemmas and that, in doing so, they take themselves to be acting in accordance with the requirements of rationality.[259] Employing the method of wide reflective equilibrium, McMahon claims that this is reason enough to accept that contribution to such cooperative schemes is justified, and that the theoretical task at hand is simply to show how the requirement to

[259] McMahon, *Collective Rationality and Collective Reasoning*, p. 15.

contribute is best understood. On his view, it is the PCR that governs the decision whether to contribute to cooperative enterprises. Yet, McMahon's approach to rational cooperation has met with severe criticism. As already emphasized, the common complaint against attempts to explain the rationality of cooperation in prisoner's dilemmas is that they are conceptually incoherent.[260] There is, however, a more general concern about McMahon's view on cooperation and the PCR. His interest is in whether the PCR can ground the obligation to obey the law. However, one can argue that McMahon's approach to the problem of political obligation is fundamentally misguided. The critic can point out that the fact that many people do obey the law has little (if any) bearing on the issue of whether there is an obligation to obey the law. Clearly, philosophical anarchists do not wish to deny that people actually do comply with the law; what they remain skeptical about is the possibility of establishing that there is an obligation to obey the law. Thus – the argument would go – approaches to political obligation which claim that the fact that most people obey the law carries justificatory weight, and that our main theoretical task is simply to uncover the reasons implicit in this practice, reverse the proper order of justification, and therefore, are objectionable. According to McMahon, the most promising strategy in responding to the challenge of philosophical anarchism is to maintain that political obligation is to be understood as grounded in the PCR. This is precisely because, on his view, what allows agents to view themselves as justified in contributing to cooperative schemes is the PCR (i.e. the PCR

[260] More specifically, the complaint is that, on the revealed preference theory, it is tautological that rational agents will defect. However, as Gerald Gaus rightly emphasizes, we should not lose sight of the fact that McMahon departs from standard utility theory in important ways, so McMahon's attempt to solve the prisoner's dilemma deserves careful consideration. For more on this point, see Gaus, *Once More Unto the Breach, My Friends, Once More: McMahon's Attempt to Solve the Paradox of the Prisoner's Dilemma*, p. 160ff. For a different line of criticism, see Gaus, *Review of Christopher McMahon's "Collective Rationality and Collective Reasoning"*.

allows agents to view prisoner's dilemma situations as having the structure of ordinary coordination problems by assigning the same payoff to unilateral defection and to the noncooperative outcome). Leaving aside the question of whether McMahon succeeds in establishing that one can have under the PCR sufficient reason to comply with the law and that the PCR can thus justify compliance with the law merely in virtue of its being the law, the problematic aspect of McMahon's approach to political obligation is that, absent certain cooperative dispositions, the PCR cannot fulfill the role that McMahon assigns to it.

Conventionalism can help to explain why an approach to political obligation such as the one advanced by McMahon is not misguided. Consider, first, the claim that it is rational to cooperate in prisoner's dilemmas. We have already seen that McMahon shares with conventionalists the idea that a promising strategy in dealing with the problem of political obligation would be to find a way of extending the treatment of ordinary coordination problems to multi-person prisoner's dilemmas. We have also seen that arguments to the effect that rationality prompts agents to cooperate in prisoner's dilemmas are typically met with suspicion by game theorists. Conventionalism can help to dispel doubts about whether the analysis of ordinary coordination problems can plausibly be extended to other game theoretic situations, such as the prisoner's dilemma, and hence, it can establish the plausibility of the claim that it is rational to cooperate in prisoner's dilemmas. Conventionalists argue at length that regularities in behavior of the sort that Lewis' analysis focuses on can emerge not only in coordination situations, but also in other game theoretic situations. Presumably, as long as the claim advanced by conventionalists is formulated in terms of regularities of behavior, it will meet with less

resistance. And we should keep in mind that conventionalists take pains to prove precisely that such regularities of behavior *can* emerge in situations in which the standard conception of rationality predicts otherwise. Drawing implications about the conception of rationality that should ultimately be adopted is a separate issue. Yet, if the conventionalist argument to the effect that a convergent pattern of behavior (and a pattern of interdependent expectations that supports it) can exist in game theoretic situations in which cooperation does not seem to be a requirement of the standard conception of rationality, what should be conceded, at the very least, is that cooperating is *not* irrational. Even if the conventionalist needs to make further assumptions (such as assuming the existence of cooperative dispositions) in order to explain why cooperating in such situations is not irrational, this should not been as troubling. If dispositions to behave cooperatively are indeed widespread, there is no reason why our theories about patterns of convergent behavior and interdependent expectations that support them should not take this fact into account.[261] We should keep in mind that the main theoretical interest of those who attempt to provide a conventionalist account of norms is not game theory *per se*. Their aim is simply to put forward a plausible account of how norms of conventions

[261] It is a further question how the emergence of cooperative dispositions is best understood. One strategy is to claim that there are good evolutionary reasons for their emergence. However, this strategy is not free of difficulties. For instance, Verbeek argues that evolutionary explanations of the predominance of cooperative dispositions are problematic because they lead us to accept a paradoxical conception of rationality. The proponents of such explanations reject the picture according to which cooperative virtues are rational in the sense of being the appropriate object of an intentional choice of the individual. According to them, it is the process of evolution that accounts for the rationality of being cooperatively virtuous. If this is case, however, it seems that instrumental rationality is in an important way self-defeating, given that it is instrumental rationality that condemns cooperative virtues as irrational in the first place. (See Verbeek, *Instrumental Rationality and Moral Philosophy: An Essay on the Virtues of Cooperation*, chapter 6.) On Verbeek's view, given that indirect justifications of cooperative dispositions are liable to the above-mentioned objection, there is no alternative but to attempt to come up with a direct justification of such dispositions. More specifically, his proposal is to give up any attempt to reconcile cooperative dispositions with the standard notion of rationality and to understand cooperative dispositions as strategic intentions. On this view, in a particular version of the prisoner's dilemma, it is rational to form such strategic intentions and to act on them. (See Verbeek, *ibid*., chapter 8.)

maintain themselves. And if it turns out that the most tenable account of why individuals conform to conventions or norms makes essential use of the idea of strategic rationality (i.e. it makes use of the idea of interdependent reasons), but at the same time implies modifying the set of assumptions that are taken for granted by game theorists when modeling behavior under highly idealized circumstances, then objecting to such an account simply on the ground that it relies on a different set of theoretical assumptions seems misplaced.

A brief point of clarification about the assumptions that the conventionalist needs to make in order to be able to show that the treatment of ordinary coordination problems can be extended to other game theoretic situations is in order here. Recall that Verbeek claims that the question of how to understand the emergence of norms is prompted by considerations of consistency. The worry is that the conventionalist may not be able to come up with an explanation of the emergence of norms consistent with the explanation of their stability. Given that the stability of norms is accounted for from the point of view of the individual, the conventionalist has to show how conventions or norms could emerge in cases that satisfy the requirements of "methodological individualism".[262] However, when he discusses the game theoretic situations in which conventions or norms could emerge, Verbeek makes certain remarks which suggest that at the end of the day conventionalism cannot satisfy the requirements of methodological individualism. More concretely, he stresses that there is nothing problematic or dilemmatic about the choice of the individual in a prisoner's dilemma, and that, under standard assumptions of rationality, it is straightforwardly rational to defect. As he puts it, the dilemmatic

[262] Verbeek, *Instrumental Rationality and Moral Philosophy: An Essay on the Virtues of Cooperation*, p. 29.

character of the game is present only if "one looks at the game from the point of view of both agents together, or if you will, from the collective point view" (*ibid*., p. 36). This is precisely why Verbeek proposes to refer to the specific version of the prisoner's dilemma in which a convention could emerge as the "prisoners' dilemma" (*ibid*., p. 35). Yet, one can point out that this seems to contradict the claim that conventionalism can satisfy the requirements of methodological individualism in explaining how conventions could emerge.

In order to answer this objection, it may be useful to take a closer look at den Hartogh's gloss on the concept of individualism. Den Hartogh points out that his version of conventionalism is committed to ontological, rather than methodological, individualism.[263] By this he means that beliefs, desires and intentions belong to individual agents. In his view, it is important to hold on to a notion of individual, as opposed to collective, agency. More specifically, he believes that we should resist the attempt to interpret mutually adjusted action by reference to a strong concept of collective intention. As he puts it, "social reality is to be understood by the way in which the beliefs, desires and intentions of individuals *depend upon* each other"[264]. Den Hartogh goes on to stress that in order to explain individuals' behavior it is not sufficient to show that it is governed by a rule, what is needed is an explanation of why the rule is followed. And arguably, any plausible explanation to this effect will have to be formulated from standpoint of the individual. However, den Hartogh argues that it is equally important to guard against interpreting individualism as atomism, and thus look for explanations of behavior which only make reference to the beliefs, desires and intentions that an individual would have if

[263] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 6.
[264] Den Hartogh, *Mutual Expectations: A Conventionalist Theory of Law*, p. 6, emphasis in original.

he were the only one possessing such mental states. As den Hartogh rightly emphasizes, we have reasons to doubt that such explanations could ever be accurate. Moreover, it should be noted that this sort of atomism is inconsistent with the strategic concepts on which conventionalism relies. The proponent of conventionalism can insist that there is no reason why the fact that conventionalism employs the idea of strategic rationality in explaining behavior should count as an argument against it. Going back to the issue that concerns us here, if the conventionalist cannot do without appealing to the idea of strategic rationality in explaining how certain norms could have emerged, there is hardly any reason why this should fuel concerns about consistency, given that, according to the conventionalist, strategic rationality is part and parcel of any plausible explanation about how such norms maintain themselves.

So far, I have argued that the fact that conventionalism relies on the assumption that people are cooperatively disposed should not be viewed as objectionable. However, one can argue that, even if it may seem that this assumption is innocuous enough when what is argued is that patterns of convergent behavior (and interdependent expectations that support them) can emerge in prisoner's dilemmas, we will nonetheless have to admit that relying on this assumption is objectionable as soon as we will take a closer look at the conventionalist attempt to justify political obligation. On the conventionalist view, the fact that individuals are cooperatively disposed plays a key role in accounting for the existence of patterns of mutual expectations, whereas the existence of such patterns plays a key role in accounting for political obligation. Thus, given that the issue of whether there are political obligations depends on whether individuals are cooperatively disposed, one can point out that assuming that individuals have cooperative dispositions means to

assume a lot. I believe, however, that conventionalism has the resources to show both that this is not an unreasonable assumption to make and that, appearances notwithstanding, the order of justification within a conventionalist account of political obligation is appropriate. We have already seen that one of the intuitions that underlies McMahon's attempt to justify political obligation by reference to the PCR is that establishing that individuals have sufficient reason to comply with authoritative directives depends on assuming that individuals are cooperatively disposed. While sharing McMahon's intuition, conventionalists take a step further in explaining why this order of justification is not inadequate. The attempt to justify the normative power of expectations is at the heart of the conventionalist enterprise. This is precisely why conventionalism holds the promise of accounting for political obligation. On this view, the existence of cooperative dispositions allows patterns of convergent behavior and patterns of mutual expectations that support them to emerge. Such patterns of mutual expectations, in their turn, generate obligations. If the conventionalist suggestion that we can move from mutual expectations to obligations can be vindicated, then the objection that questions the order of justification within a conventionalist account of political obligation is met.

Chapter 6
Conclusion

My main purpose in this dissertation has been to provide an answer to the question whether interdependent reasons for action can help provide a solution to the problem of political obligation. I have begun by examining the interaction between autonomy, authority and rationality, focusing on the potential conflict between the personal autonomy of the moral agent and political authority. I have attempted to substantiate a reading of the challenge put forward philosophical anarchists, according to which what allows autonomy to come in contradiction with authority is the fact that the former is allied with rationality.

This reading led to me to investigate one of the most elaborate recent attempts to respond to the charge that submitting to the directives of authority is irrational. We have seen that, according to McMahon, even though individuals may not a have reason to comply with the directives of authority as long as they consider the matter from the standpoint of individual rationality, they may nevertheless have a reason to comply with them if they consider the matter from the standpoint of collective rationality. My aim was to explore the strengths and limitations of a model that attempts to provide the normative foundations of political obligation by employing the principle of collective rationality. I have pointed out that the principle of collective rationality plainly dictates a greater level of compliance with the law than the principle of individual rationality in conjunction with an individual's moral values dictate. However, the worry is that, while the principle of collective rationality may ultimately fail to ground a suitably general obligation to obey

the law, the considerations that are invoked to establish in the first place the need to appeal to a principle that would give one a reason to obey the law over and above what moral principles that work within the framework of individual rationality give one reason to do, might render implausible any attempt to account for political obligation by reference to such substantive moral principles. I have attempted to dispel this worry by arguing that, while McMahon is right in claiming that that there is something deeply troubling about the suggestion that one has to manifest a fundamental disposition to make concessions when what is at stake is the realization of one's moral concerns, we can resist the conclusion that considerations of fairness cannot appropriately be invoked when attempting to justify political obligation. More concretely, I have argued that McMahon's argument trades on an equivocation between different types of situations in which considerations of fairness can be invoked, and that the claim that fairness cannot legitimately be invoked in justifying political obligations ultimately depends on the claim that it is not possible that the choice of a cooperative scheme is regarded by all as guided by reason. Furthermore, I have suggested that the resources provided by a conventionalist account may help take a significant step in the direction of rehabilitating the idea that the reason to comply with cooperative schemes can be accounted for by appeal to substantive moral considerations.

Conventionalism constitutes an interesting venue for the study of political obligation given that, on this view, each participant to a convention has a reason to conform to it as long as everyone else conforms to it. According to the Lewisian inspired version of conventionalism that I have focused on, collectively stable cooperative strategies manage to emerge and maintain themselves precisely because they serve

individuals' mutual interests. I have pointed out that conventionalists share with McMahon not only the idea that political obligation, if it exists, can only be accounted for in terms of interdependent reasons for action, but also the idea that a promising strategy in dealing with the problem of political obligation would be to show that the treatment of ordinary coordination problems can be extended to other game theoretic situations, such as the prisoner's dilemma. I have attempted to show that the suggestion according to which certain basic aspects of the Lewisian conventionalist model are also present in situations of considerable conflict of interests is on the right track. This allows those who attempt to provide a conventionalist account of authority and political obligation to answer the objection that such attempts are misguided because the conventionalist analysis cannot properly be applied to situations of significant conflict of interests. As I have argued, dropping the condition which requires similarity in preference for a convention to exist is not a dubious move.

On the conventionalist view, political obligation is justified by reference to a pattern of mutual expectations. The version of conventionalism that I have focused on makes crucial use of the idea of cooperative dispositions. I have attempted to show that conventionalists are right in claiming that assumptions about cooperative dispositions are not *ad hoc*, and therefore, conventionalism can make sense of the idea that patterns of mutual expectations are not immune to moral criticism. Furthermore, I have argued that conventionalism has the resources to answer the most compelling objections against conventionalist accounts of authority and political obligation. Critics often claim that conventionalism misrepresents the categorical nature of obligation. However, I have argued that it would wrong to think that conventional norms have only hypothetical

authority. Moreover, I have emphasized that one can insist that the fact that conventionalism cannot account for the alleged unconditional nature of political obligation, should be viewed as a strength rather than a weakness of the theory, since one can contend that political obligation, if it exists, can only be a conditional and selective duty. Also, I have argued that conventionalism has the resources to show that one can have an independent moral reason to comply with authoritative directives.

Appendix
The plural subject theory of political obligations

1 Margaret Gilbert's plural subject theory

One can distinguish between two types of philosophical accounts of political obligation.[265] On the one hand, theories of acquired obligation attempt to prove that the obligation to contribute to the maintenance of the state is based on consent or, according to the principle of fairness, on willingly accepting the benefits that result from cooperation. On the other hand, theories of natural duty aim to show that the duty to support the political institutions of a *just* state is not dependent on contingent factors such as consent or receipt of benefits.

Theories that belong to the first class are typically *voluntarist* theories. More concretely, they claim that political obligations can only arise from individuals' *voluntary* choices to subject themselves to political authority or to participate in ongoing cooperative social schemes. However, these theories cannot yield the conclusion that people do in fact have political obligations unless the claim that actual political societies are in important respects similar to voluntary associations is a descriptively accurate claim. The objection that individuals did not actually choose to participate or to become members of their societies leaves the proponents of theories of acquired obligation with several ways of arguing. On the one hand, taking the consent approach, a possible strategy is to turn actual (or tacit) consent into hypothetical consent, and to argue that

[265] For this distinction, see Jeremy Waldron, *Special Ties and Natural Duties*, p. 3.

people would presumably consent to a just system if they were asked. On the other hand, one can adopt the fair play approach and argue that an individual's receipt of benefits from a cooperative scheme amounts to that individual's being treated justly by the scheme, and this, in turn, generates her obligations.[266]

However, the proponent of the theory of acquired obligation might take a different stance in the dispute concerning the voluntary or nonvoluntary character of membership. Instead of arguing that, appearances notwithstanding, the voluntariness condition does hold, and thus, it entails that individuals have political obligations, one can accept that group membership is constituted by acts that are not always fully voluntary, while holding that this is not a troubling conclusion in what political obligation is concerned. This is the position defended in several papers by Margaret Gilbert. Her theory, labeled by John Simmons as a "nonvoluntarist contract theory", allows that individuals can accrue political obligations by acts which are not necessarily fully voluntary.[267]

In what follows, I will proceed to explain in what sense Gilbert's theory of political obligation is a *nonvoluntarist* theory. In *Reconsidering the 'Actual Contract' Theory of Political Obligation*, Gilbert emphasizes that the "actual contract" theory of political obligation, according to which individuals' obligations to uphold the political institutions of their country are founded in the agreement that makes a particular country their country, has to face two basic objections.[268] First, the "no agreement objection"

[266] Waldron argues that although these last two views are formulated in terms of acquired obligation, they are pushed in the direction of natural duty accounts. This is because the moral requirement of obedience is not contingent on anything that the individual has said or done, but it is predicated on the claim that the laws of the state embody demands of *justice*. See *Special Ties and Natural Duties*, p. 4.
[267] John Simmons, *Associative political obligations*, p. 255.
[268] Gilbert, *Reconsidering the 'Actual Contract' Theory of Political Obligation*, pp. 238–241.

claims that no relevant agreement can be plausibly said to exist in most actual political societies, regardless of whether agreement is conceived as "tacit" or "implicit".

Gilbert construes her plural subject theory in response to the above-mentioned objection, which she claims is inconclusive. Her argument runs as follows. Certain shared activities presuppose the existence of a plural subject. A plural subject is in place whenever two or more people are jointly committed to doing something as a body. Gilbert's idea of commitment is rather loose, it can range from explicit agreement to a sort of informal or tacit agreement, or to a vague mutual understanding. However, the necessary and sufficient condition for a joint commitment to come into being is that "the relevant parties mutually express their readiness to be so committed, in conditions of common knowledge".[269] The function of joint commitments is "to establish a set of obligations and entitlements between individual persons to establish a special 'tie' or 'bond' between them".[270] The fact that obligations and entitlements are inherent in any joint commitment, becomes manifest if one considers one of the simplest cases of shared activities, that of two people going for a walk together.[271] Following Gilbert, it is plausible to suggest that the concept of obligation applies to this case, since if one person would suddenly change her mind and turn away we would consider this type of behavior quite odd. Thus, the least one can assert about the case of two people walking together is that they have the obligation not to leave without saying a word. This, in turn, amounts to the claim that each party has an obligation to conform to the joint commitment, and that neither of them is in a position to rescind it unilaterally.

[269] Gilbert, *Group Membership and Political Obligation*, p. 123.
[270] Gilbert, *Group Membership and Political Obligation*, p. 124.
[271] Gilbert, *Group Membership and Political Obligation*, p. 123.

A second step in the argument is to interpret political obligation in terms of membership in a plural subject with an appropriate underlying joint commitment. Following Gilbert line of argument, "social groups are plural subjects; plural subjects are constituted by joint commitments which immediately generate obligations".[272] It should be noticed that this second part of the argument crucially depends on the fact that the plural subject theory is not built on *voluntarist* assumption. The fact that one can enter a joint commitment with a *minimum of voluntariness*, that the expression of the readiness to be jointly committed is sufficient and that explicit agreement is not required, is especially relevant for the case of political obligation. The less is required for a joint commitment to be in place, the more easily the commitment that yields political obligations can be said to be a matter of "common knowledge". Thus, Gilbert's intention is to make her theory immune to the objection that the individuals' taking themselves to be party to a joint commitment of the relevant sort is not well-founded. She argues that "the widespread observed use of such phrases as 'our government' and 'our country', alongside any relevant behavior, would itself appear to provide a basis for common knowledge in a population that a substantial portion of its members have openly expressed their willingness jointly to commit in the relevant way".[273]

What follows from the above-mentioned considerations is that compared to actual contract theory, Gilbert' plural subject theory of political obligation presents the advantage of requiring a *minimum of voluntariness*.[274] This proves to be an advantage given that agreements which are supposed to generate political obligations can more

[272] Gilbert, *Group Membership and Political Obligation*, p. 126.
[273] Gilbert, *Group Membership and Political Obligation*, p. 128.
[274] Gilbert suggests that her theory would be best referred to as *intentionalist* rather than *voluntarist* theory. *Reconsidering the 'Actual Contract' Theory of Political Obligation*, p. 254.

plausibly said to exist in actual political circumstances once they are interpreted in terms of joint commitments. This is not only because, as mentioned above, the expression of the readiness to be committed is sufficient to give rise to obligation, but also for a further reason, namely that joint commitments need not involve any *datable* act of commitment. Joint commitment might be established at a particular moment in time, but might also arise as a result of a gradual process, as in the case of two academic colleagues, who realize after several ad hoc arrangements that they are jointly committed to have dinner together after departmental meetings.[275]

Now, let me turn to the second standard objection to the "actual contract" theory that Gilbert's account of political obligation attempts to meet. The "not morally binding objection" concedes the existence of agreements of the relevant sort. However, it claims that either the circumstances or the content of such agreements would prevent them from being morally binding. First, supposing that in practice the alternative to the refusal to agree would be highly costly (i.e. emigration, imprisonment or social ostracism), an agreement entered in such circumstances would amount to a coerced agreement, and consequently, would result in its not being morally binding. Secondly, given that certain laws of certain states are unjust or immoral, the agreement to uphold the political institutions of such states would similarly be prevented from being morally binding. In response to this objection, Gilbert argues that joint commitments need *not* be *fully voluntary*, in the sense that entering them and being obligated by them is not precluded by coercive circumstances or immoral content.[276] She claims that if agreements are to be

[275] Gilbert, *Reconsidering the 'Actual Contract' Theory of Political Obligation*, p. 243.
[276] Gilbert, *Reconsidering the 'Actual Contract' Theory of Political Obligation*, pp. 248–249.

conceived as being underlined by joint commitments, then obligations of joint commitment will be in place whenever an agreement is made.

2 A critique

In what follows, I will attempt to offer a critical analysis of Gilbert's plural subject theory of political obligation. My analysis will attempt to answer three questions. First, I will focus on the question whether Gilbert is right in asserting that joint commitments do establish obligations and entitlements. The second question that I will try to answer is whether obligations and entitlements derive from joint commitments irrespective of any other considerations. More concretely, I will consider the question whether there are sufficient reasons for accepting that being obligated by a joint commitment is not precluded by coercive circumstances or coercive content. Finally, I will try to assess the difficulties involved in trying to apply Gilbert's general account of joint commitments to the political sphere in a way that will yield political obligations.

Let me start by trying to answer the first question. On what grounds can we assert that the parties to a joint committed are under an obligation to conform to it? As M. E. Bratman suggests, "the relation of shared intention to mutual obligation is to be determined by identifying relevant principles of obligation".[277] Following further Bratman's suggestion, the most plausible candidate for the principle that could ground the obligation to act as one has indicated is the principle that ties obligation to the purposive creation of expectations. Although Gilbert does not explicitly say that the

[277] Bratman, *Shared Intention and Mutual Obligation*, p.138.

obligation to conform to joint commitments rises from the fact that harm is done by frustrating expectations, this interpretation is compatible with her account. Consider for instance her example of the academic colleagues who realize after several ad hoc meetings that they are jointly committed to leave together after departmental meetings. Gilbert contends that each of them has the obligation not to interrupt this already established practice without a special justification, without reaching a mutual accord concerning a possible non-conformity. She seems to suggest that doing otherwise would frustrate the expectations of one of the persons involved about the other's conduct.

However, we can further ask whether the requirement not to frustrate others' expectations is the exclusive ground on which conformity to joint commitments can be prescribed. Are there any other principles compatible with Gilbert's account of joint commitments that would yield obligations of conformity? There are at least two reasons why this is not the case.

First, it is doubtful whether a mental attitude alone can be the source of some obligation. To paraphrase Kent Greenawalt, if a given attitude represents simply the recognition of all the valid arguments for obeying, say a moral law, then the attitude itself adds nothing to the force of these arguments.[278] If on the other hand, the attitude reflects a mistaken assessment of that given moral law, then the attitude does not constitute an independent obligation to obey the law. The claim that a mental attitude is enough to underlie an obligation seems stronger in the case when the attitude is freely chosen, i.e. when the person does not think she is morally compelled to adopt that attitude. However,

[278] The question that Kent Greenawalt actually discusses is whether a mental attitude can be the source of the obligation to obey the law. However, his arguments can be generalized in such a way as to contend that a mental attitude is generally not sufficient to underlie obligations. See *Conflicts of Law and Morality*, p. 75.

Greenawalt argues, even considering this case, we will arrive at the same conclusions. He acknowledges that "there may be some virtue, perhaps moral virtue, in sticking to certain choices once they have been made even if choices could have been made differently and even if no one else is relying on them".[279] On the other hand, he claims that making an agent blameworthy for the fact that she did not carry out her freely chosen uncommunicated commitments is not a tenable position. The least that can be asserted is that the moral ought behind these types of commitments is *much weaker* than the one deriving from communicated commitments and mutual expectations. Greenawalt's argument is relevant here to the extent that it reinforces the idea that the ground for the obligations of joint commitment must have to do with something that is "common knowledge" between the parties, that is "communicated" in one way or another, and that can be the basis on which mutual expectations are built.

A second reason why it is most plausible to assume that the creation of expectations grounds the obligations of joint commitment is that Gilbert's theory is intended to cover a very wide range of cases. Taking for instance the case of marriage, when deciding whether to violate its rules or not, the relevant considerations might include not only those that concern the expectations of the other party, but also the value that one attaches to the institution of marriage. However, it seems rather obvious that in the simplest cases of joint commitments considerations that assign value to a certain practice do not enter the picture. It would be absurd, for instance, to talk about the "value of walking". This seems to be perfectly compatible with Gilbert's account, since she claims that obligations of joint commitment, as opposed to other moral obligations, are

[279] Kent Greenawalt, *Conflicts of Law and Morality*, p. 76.

obligations due to particular persons.[280] However, this point might be easily overlooked when applying the general account of joint commitments to the problem of political obligations. What is important is that arguments that refer to the creation of expectations do not collapse into the type of arguments aiming to show that there is an obligation to go along with the existing arrangements based on the idea that behind these arrangements there is a certain value that is worth being promoted. Moreover, since Gilbert's theory has to account for the simplest cases of joint commitment, like the one of two people going for a walk together, we have to accept that the ground for obligation is the one referring to the creation of expectations, given that no stronger tie can plausibly said to exist between two people who barely know each other.

Now let me clarify Gilbert's position on whether an explicit, or intentional, creation of expectations is needed in order to assert that one is under the obligation to act as expected, or whether a misunderstanding on the other part is sufficient to yield the obligation. Gilbert seems to allow for both views. On the one hand, she holds that "one's being obligated depends in part on one's intentionally obligating oneself".[281] This claim is consistent with the idea that the necessary condition for entering a joint commitment is that the relevant parties express their readiness to be committed, in conditions of common knowledge. On the other hand, since she claims that joint commitments need not involve any datable act of commitment, but can simply grow up somehow, we can imagine a case when someone takes herself to be party to joint commitment by misinterpreting the other person's behavior. What matters is not whether Gilbert would hold that in this particular case a joint commitment is in place since, presumably, she would claim that the person

[280] Gilbert, *Reconsidering the 'Actual Contract' Theory of Political Obligation*, p. 246.
[281] Gilbert, *Reconsidering the 'Actual Contract' Theory of Political Obligation*, p. 254.

who does not feel herself party to a joint commitment would have obligations similar to the ones that a joint commitment yields. More concretely, an argument compatible with her ideas would be the following. The person ought to act as she is expected in order not frustrate the other's expectations, at least as long as only a small inconvenience is involved. If great inconvenience would be caused by trying not to hurt the other's feelings, then the person would have only the obligation to settle the misunderstanding in order to prevent forming future mistaken expectations.

To sum up, we can say that, according to Gilbert, an intentional creation of expectations is needed in order to hold someone to be under the obligation to act as expected. However, given the value she assigns to the moral principle that harm is done by frustrating expectations, it is consistent with her views to the claim that, as long as only a small inconvenience is caused, one ought not to frustrate others' expectations. Nevertheless, what is especially relevant here is that in neither of the two cases is one obligated to act *unconditionally* in conformity with others' expectations. On the one hand, in the case when one has intentionally obligated herself, the obligation is in place only as long as the joint commitment is in place. Even if a person cannot unilaterally rescind the commitment, it seems enough that one party openly expresses her whish not to be further committed. Although joint commitments provide a genuine basis for obligations, Gilbert acknowledges there are also other considerations that are relevant when deciding what one *ought* to do. On the other hand, the requirement to act as expected will be even less stringent when considering the case of a person who did not intentionally created the expectations that she will act in certain manner. One cannot

plausibly defend a theory that would entail that one ought to marry a mad person because she expects this to happen.

3 Interpreting political obligation in terms of joint commitment

All these considerations lead to the conclusion that there are difficulties in applying Gilbert's account of joint commitments to the political sphere. Most individuals are born in political societies and even before reaching maturity they already participate in the ongoing schemes of social cooperation. Thus, without realizing, individuals are part to joint commitments. However, they cannot under *any* circumstances, rescind their commitments, nor can they claim that until a particular moment (say the age of maturity) their behavior was not an intentional, or conscious creation of expectations. This proves that Gilbert's general account of the obligations deriving from joint commitments is not a good analogy for political obligation. Following Gilbert's own account, the obligations of joint commitment do not provide *conclusive* reasons for action, while the commands of authority are supposed, on a most plausible analysis, to provide such reasons.[282]

There are also other considerations that plead against interpreting political obligation in terms of joint commitment. Once we concede that frustrating expectations is the exclusive harm involved in violating the obligations of joint commitment, we might ask whether there are any cases of shared intention that fail to involve mutual obligation.

[282] The analysis that I have in mind is one that regards authoritative commands as offering exclusionary reasons for action, and is put forward by Joseph Raz in several of his writings.

Bratman argues that an exchange of disavowals results in blocking the associated obligations. He claims that:

> […] if I indicate that I reserve the right to change mind at will, I thereby indicate that, though I am trying to lead you to expect that I will sing my part, I am not trying to lead you to expect that I will sing my part unless you consent to my not singing.[283]

The idea is that the parties to a joint commitment could reserve themselves the right to rescind it unilaterally at any given moment in time if it becomes manifest that none of the parties' expectations will frustrated by doing so. Gilbert claims that her account can accommodate Bratman' suggestion:

> Any plausible 'reservation of right' of this kind is likely to be something explicitly agreed by the parties. It could, of course, initiate a convention so that it became unnecessary in particular cases explicitly to bring up the side understanding in question.[284]

However, in what the social or political sphere is concerned, *no one is in the position to rescind the commitment unilaterally* since there are no clearly identifiable parties that could explicitly make such an agreement. This proves again that the analogy between the obligations that derive from joint commitments and political obligations is not a good analogy. As I noticed above, the obligations of joint commitment are due to *particular* persons. As long as it is clear who would be harmed by not conforming to the joint commitment, harm can be avoided even if the joint commitment is rescinded, since in some circumstances it can be rescinded by both parties. However, if these obligations are due to *all* the individuals that participate in the social cooperative scheme, then the

[283] Bratman, *Shared Intention and Mutual Obligation*, pp. 137–138.
[284] Gilbert, *Reconsidering the 'Actual Contract' Theory of Political Obligation*, p. 245.

commitment cannot be rescinded under *any* circumstances, and those who refuse to cooperate cannot avoid doing harm.

Furthermore, one might ask whether the requirement not to frustrate others' expectations has the same normative consequences regardless of whether one considers direct, personal activities or indirect, impersonal shared activities. Simmons's distinction between *reasonable expectations* and *entitlements* suggests that this is not the case.[285] According to Simmons, the relationships that typically hold between fellow subjects in large-scale political communities resemble the indirect, impersonal relationships which create reasonable expectations, but does not imply any entitlement. To illustrate this point one can compare the direct and personal relationships, laden with tacit commitments, which obtain between friends who play bridge on regular basis, with the indirect and impersonal relationship between Kant and the housewives of Konigsberg, whose reasonable expectations are to be able to set their clocks by Kant's walks. While the former relationships could be said to ground a certain kind of entitlement that some of the friends in case expect the others to continue showing up for the bridge game, the housewives from Konigsberg are not entitled to demand that Kant fulfills their expectations. Given that the relationships between members of large-scale political communities resembles closely the latter case, the commitments that holds within these communities is too weak to ground political obligation.

One can grant that Gilbert is right in suggesting that a personal and direct relationship (such as friendship) is not needed in order for the obligations of joint commitments to be in place. Suppose that two colleagues who barely know each other end up by going for a walk together one day, when they discover that their way home

[285] Simmons, *Associative Political Obligations*, pp. 257–259.

partly coincides. As already suggested, in such a case we would consider it odd if one of them would leave without saying a word. However, we should observe, with Simmons, that there is an asymmetry between cases of direct, personal shared activities and indirect, impersonal ones. To take the extremes, we can compare the "walking together" case with that of social cooperation. In the former case, the requirement not to frustrate one's expectations is supposed to yield a weak obligation (i.e. a simple justification is enough to abstain from discharging it). By contrast, in the latter case the same requirement is supposed to yield a strong obligation (which can be overridden by few, if any, considerations). Still, the expression of the readiness to be committed which might often look like an agreement (or consent) seems to be necessarily present in the first case, but not in the second. Therefore, we can conclude with Simmons, that in the second case, as opposed to the first one, we cannot properly say that there is an entitlement or obligation.

One can reply that the obligation to go along with social or political arrangements is stronger (a simple justification will not be enough to abstain from discharging it) given that, compared to other cases of shared activities, more harm can be done by non-conformity. However, it is equally true that what the individual is required to do is not something that involves minimal costs, as going for a walk, but planning her entire life in light of these considerations. Consequently, it seems plausible to assume that a *stronger notion of commitment* is needed in order to explain individuals' political obligations. Nevertheless, Gilbert cannot make this move since, as Simmons rightly points out "insisting on the stronger notion of commitment she needs would in effect involve

reasserting the voluntarist picture of political society whose rejection partly motivated the project in the first place".[286]

Finally, let me consider briefly Gilbert's case for the idea that coercive circumstances or immoral content do not prevent an agreement from being binding. What is especially relevant in this context is not whether Gilbert's position is, as such, a tenable position, but rather the normative consequences of her claims for the problem of political obligation. In *Agreements, Coercion and Obligation*, Gilbert considers the following gunman case: Emily's father forces Ben to agree to marry her. Gilbert acknowledges that Ben's coercer is not morally entitled to his performance, and in the case when Ben does not conform to the agreement, he has no basis for complaint against Ben, given the way the agreement was brought about. However, the fact that Ben was coerced to agree to marry Emily is not an independent consideration against marrying Emily. As Gilbert argues:

> […] if we define a coerced agreement in terms of a threat of violence that in fact motivated a person to agree, the definition does not tell us that there are independent reasons against keeping the agreement once it is made.[287]

Considering the case of coerced agreements, the first point that can be made is that it is quite odd to claim that the obligation to conform to the joint commitment derives from the requirement not to frustrate others' expectation. As opposed to other cases of joint commitment, we cannot claim that greater harm is involved by frustrating the expectations that the coercer has than by performing an act to which one agreed in coercive circumstances. Gilbert might be right in claiming that a coerced agreement might simply introduce an extra motivational force into a situation where there are no

[286] Simmons, *Associative Political Obligations*, p. 259.
[287] Gilbert, *Agreements, Coercion and Obligation*, p. 703.

independent reasons *against* keeping the agreement. However, we can also doubt whether the agreement does in fact provide an independent reason *for* acting in conformity with its prescriptions. Thus, we can question the claim that a coerced agreement provides a genuine basis for obligation.

More importantly, Gilbert stresses that the party who was coerced to enter an agreement is morally entitled to renege it. For the purpose of applying this argument to the case of political obligation it does not matter whether we accept that there is no actual obligation or that there is an obligation that is overridden by more compelling considerations. The relevant point is that this argument will not yield the desired conclusion when applied to political obligations. By claiming that the parties who are coerced to enter an agreement are morally entitled to renege it, one cannot account for the *general* character of political obligation, nor for the fact that authoritative demands provide *conclusive* reasons for action.[288]

In conclusion, we can say that, although joint commitments can establish obligations, these obligations do not derive from joint commitments irrespective of any other considerations. Moreover, one can claim that there are serious difficulties involved in the attempt to apply the plural subject theory to the case of political obligation. Consequently, we can conclude that Gilbert's plural subject theory does not provide sufficient reasons for asserting that people do in fact have political obligations.

[288] Note that there are at least two considerations that recommend the conventionalist account of political obligation as a more tenable account than the one advanced by Gilbert. First, conventionalists make room for the idea that mutual expectations are not immune to moral criticism. Thus, on a conventionalist account, mutual expectations assume an obligating character to the extent that they are legitimate. Second, even though conventionalism fails to justify the existence of a general political obligation, it has the resources to show why the quest for such a justification is misguided. If the conventionalist analysis of norms turns out to be coherent and plausible, then a case can be made that political obligation, if it exists, can only be a conditional and selective obligation.

Bibliography

Aumann, Robert. 1987. Game Theory. In J. Eatwell, M. Milgate and P. Newman (eds.), *The New Palgrave: A Dictionary of Economics, Volume 12*. London: Macmillan Press, 460–482.

Axelrod, Robert. 1981. The Emergence of Cooperation Among Egoists. *American Political Science Review* 75: 306–318.

—— 1984. *The Evolution of Cooperation*. New York: Basic Books.

Axelrod, Robert and Hamilton, W.D. 1981. *The Evolution of Cooperation*, *Science* 211: 1390–1396.

Arneson, Richard J. 1982. The Principle of Fairness and Free-Rider Problems. *Ethics* 92: 616–633.

Baier, Kurt. 1958. *The Moral Point of View: A Rational Basis of Ethics*. Ithaca, NY: Cornell University Press.

Binmore, Ken. 1993. *Game Theory and the Social Contract, Volume 1: Playing Fair*. Cambridge, Mass.: MIT Press.

—— 1999. Review of Martin Hollis' "Trust within Reason". *Economic Journal* 109: 211–212.

Bird, Colin. 2011. Comment on Gerald Gaus's "The Order of Public Reason". Online at URL= < http://publicreason.net/2011/03/31/opr-vii19-coordinating-on-a-morality/>.

Blackburn, Simon. 1984. *Spreading the Word*. Oxford: Oxford University Press.

—— 1998. *Ruling Passions*. Oxford: Oxford University Press.

Bratman, Michael. 1998. Toxin, Temptation and the Stability of Intention. In Jules Coleman and Christopher Morris (eds.), *Rational Commitment and Social Justice: Essays for Gregory Kavka*. Cambridge: Cambridge University Press, 59–83.

—— 1999. Shared Intention and Mutual Obligation. In Michael Bratman, *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge: Cambridge University Press, 130–141.

Brink, David. 2003. *Perfectionism and the Common Good: Themes in the Philosophy of T. H. Green*. Oxford: Oxford University Press.

Burge, Tyler. 1975. On Knowledge and Convention. *The Philosophical Review* 84: 249–255.

Buss, Sarah. 2008. Personal Autonomy. *Stanford Encyclopedia of Philosophy*. Edward N. Zalta (ed). URL http://plato.stanford.edu/archives/fall2008/entries/personal-autonomy.

Cubitt, Robin and Sugden, Robert. 2003. Common Knowledge, Salience and Convention: A Reconstruction of David Lewis' Game Theory. *Economics and Philosophy* 19: 175–210.

Dagger, Richard. 2000. Membership, Fair Play, and Political Obligation. *Political Studies* 48: 104–117.

Dickson, Julie. 2007. Is the Rule of Recognition Really a Conventional Rule? *Oxford Journal of Legal Studies* 27: 373–402.

Dworkin, Gerald. 1988. *The Theory and Practice of Autonomy*. Cambridge: Cambridge University Press.

—— 1993. Autonomy. In Robert E. Goodin and Philip Pettit (eds.), *A Companion to Contemporary Political Philosophy*. Oxford: Blackwell, 443–451.

Estlund, David. 1993. Making Truth Safe for Democracy. In David Copp, Jean Hampton and John E. Roemer (eds.), *The Idea of Democracy*. Cambridge: Cambridge University Press, 71–100.

Feinberg, Joel. 1986. *Harm to Self*, Volume 3 of *The Moral Limits of Criminal Law*. Oxford: Oxford University Press.

Finnis, John. 1984. The Authority of Law in the Predicament of Contemporary Social Theory. *Journal of Law, Ethics and Public Policy* 1: 115–137.

Fiss, Owen M. 1985. Conventionalism. *Southern California Law Review* 58: 177–197.

Flathman, Richard E. 1972. *Political Obligation*. New York: Atheneum.

Frankfurt, Harry. 1973. The Anarchism of Robert Paul Wolff. *Political Theory* 1: 405–414.

Friedman, James W. 1986. *Game Theory with Applications to Economics*. New York: Oxford University Press.

Friedman, R. B. 1990. On the Concept of Authority in Political Philosophy. In Joseph Raz (ed.), *Authority*. New York: New York University Press, 56–91.

Gaus, Gerald. 2002. Review of Christopher McMahon's "Collective Rationality and Collective Reasoning". *Notre Dame Philosophical Reviews*. URL http://cfweb-prod.nd.edu/philo_reviews/review.cfm?id=1145

—— 2003. Once More unto the Breach, My Friends, Once More: McMahon's Attempts to Solve the Paradox of the Prisoner's Dilemma. Philosophical Studies 116: 159–170.

—— 2008. Reasonable Utility Functions and Playing the Cooperative Way. *Critical Review of International and Social Philosophy* 11: 215–234.

—— 2011. *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*. Cambridge: Cambridge University Press.

Gauthier, David. 1963. *Practical Reasoning: The Structure and Foundation of Prudential and Moral Arguments and their Exemplification in Discourse*. Oxford: Clarendon Press.

—— 1986. *Morals by Agreement*. Oxford: Clarendon Press.

Gilbert, Margaret. 1989*a*. *On Social Facts*. London: Routledge.

—— 1989*b*. Rationality and Salience. *Philosophical Studies* 57: 61–78.

—— 1993a. Group Membership and Political Obligation. *Monist*: 76: 119–131.

—— 1993*b*. Agreements, Coercion and Obligation. *Ethics* 103: 679–706.

—— 1999. Reconsidering the 'Actual Contract' Theory of Political Obligation. *Ethics* 109: 236–260.

Gilboa, Itzhak. 2010. *Rational Choice*. Cambridge, Mass.: MIT Press.

Godwin, William. 1976. *Enquiry Concerning Political Justice*. Harmondsworth: Penguin Books.

Graham, Keith. 1982. Democracy and the Autonomous Agent. In Keith Graham (ed.), *Contemporary Political Philosophy: Radical Studies*. Cambridge: Cambridge University Press, 113–138.

—— 1986. The Battle of Democracy: Conflict, Consensus and the Individual. Totowa, NJ: Barnes & Noble.

—— 2002. *Practical Reasoning in a Social World: How We Act Together*. Cambridge: Cambridge University Press.

Green, Leslie. 1990. *The Authority of the State*. Oxford: Clarendon Press.

—— 1999. Positivism and Conventionalism. *Canadian Journal of Law and Jurisprudence* 12: 35–52.

—— 2003. Strategy and Fundamental Legal Rules. *American Philosophical Association Newsletter on Law and Philosophy* 3: 69–74.

Green, Thomas Hill. 1921. *Lectures on the Principles of Political Obligation*. London: Longman's, Green & Co.

Greenawalt, Kent. 1989. *Conflicts of Law and Morality*. Oxford: Oxford University Press.

Hampton, Jean. 1998. *The Authority of Reason*. Cambridge: Cambridge University Press.

Hart, H. L. A. 1955. Are There Any Natural Rights? *Philosophical Review* 64: 175–191.

—— 1961. *The Concept of Law*. Oxford: Clarendon Press.

—— 1982. *Essays on Bentham: Jurisprudence and Political Theory*. Oxford: Clarendon Press.

Hartogh, Govert den. 1993. Rehabilitating Legal Conventionalism. *Law and Philosophy* 12: 233–247.

—— 2002. *Mutual Expectations: A Conventionalist Theory of Law*. New York: Kluwer Law International.

—— 2004. The Authority of Intention. *Ethics* 115: 6–34.

Horton, John. 1992. *Political Obligation*. London: Macmillan.

Hume, David. 2001. *A Treatise of Human Nature*. Oxford: Oxford University Press.

Kavka, Gregory. 1983. The Toxin Puzzle. *Analysis* 43: 33–36.

Klosko, George. 1992. *The Principle of Fairness and Political Obligation*. Lanham, MD: Rowman & Littlefield Publishers.

—— 1998. Fixed Content of Political Obligations. *Political Studies* 46: 53–67.

Kramer, Matthew H. 1999. Requirements, Reasons and Raz: Legal Positivism and Legal Duties. *Ethics* 109: 375-407.

Kreps, David M. 1990. *A Course in Microeconomic Theory*. Princeton, NJ: Princeton University Press.

Lagerspetz, Eerik. 1989. *A Conventionalist Theory of Institutions*. Helsinki: Acta Philosophica Fennica, Vol. 44.

—— 1995. *The Opposite Mirrors: An Essay on the Conventionalist Theory of Institutions*. Dordrecht: Kluwer Academic Publishers.

Lewis, David. 1969. *Convention: A Philosophical Study*. Cambridge, MA: Harvard University Press.

—— 1975. Language and Languages. In K. Gunderson, *Language, Mind and Knowledge: Minnesota Studies in the Philosophy of* Science. Minneapolis: University of Minnesota Press, 3–35.

Lukes, Steven. 1979. Power and Authority. In Tom Bottomore and Robert Nisbet (eds.), *A History of Sociological Analysis*. London: Heinemann, 633–676.

—— 1990. *Perspectives on Authority*. In Joseph Raz (ed.), *Authority*. New York: New York University Press, 203–217.

Marmor, Andrei. 2009. *Social Conventions: From Language to Law*. Princeton and Oxford: Princeton University Press.

McDermott, Daniel. 2004. Fair-Play Obligations. *Political Studies* 52: 216–232.

McMahon, Christopher. 1987. Autonomy and Authority. *Philosophy and Public Affairs* 16: 303–328.

—— 1994. *Authority and Democracy: A General Theory of Government and Management*. Princeton, NJ: Princeton University Press.

—— 2001. *Collective Rationality and Collective Reasoning*. Cambridge: Cambridge University Press.

—— 2003. Reply to Gaus, Richardson and Weber. *Philosophical Studies* 116: 197–213.

—— 2005. Shared Agency and Rational Cooperation. *Noûs* 39: 284–308.

—— 2009. *Reasonable Disagreement: A Theory of Political Morality*. Cambridge: Cambridge University Press.

Miller, Seumas. 1991. Co-ordination, Salience and Rationality. *The Southern Journal of Philosophy* 29: 359–370.

—— 1992. On Convention. *Australasian Journal of Philosophy* 70: 435–444.

—— 2001. *Social Action: A Teleological Account*. Cambridge: Cambridge University Press.

Morris, Christopher W. 1996. Well-Being, Reasons, and the Politics of Law. *Ethics* 106: 817–833.

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

Parsons, Stephen D. 2005. Fair-Play Obligations: A critical Note on Free Riding. *Political Studies* 53: 641–649.

Pateman, Carole. 1973. Political Obligation and Conceptual Analysis. *Political Studies* 21: 199–218.

Perry, Stephen R. 1989. Second-Order Reasons, Uncertainty and Legal Theory. *Southern California Law Review* 62: 913–994.

Pettit, Philip. 1988. The Prisoner's Dilemma is an Unexploitable Newcomb Problem. *Synthese* 76: 123–134.

Pitkin, Hanna. 1966. Obligation and Consent – II. *American Political Science Review* 60: 39–52.

Postema, Gerald. 1982. Coordination and Convention at the Foundations of Law. *Journal of Legal Studies* 11: 165–203.

Poundstone, William. 1993. *Prisoner's Dilemma*. New York: First Anchor Books.

Rawls, John. 1964. Legal Obligation and the Duty of Fair Play. In S. Hook (ed.), *Law and Philosophy*. New York: New York University Press, 3–18.

—— 1971. *A Theory of Justice*. Cambridge, MA: Harvard University Press.

—— 1996. *Political Liberalism* (2nd edition). New York: Columbia University Press.

Raz, Joseph. 1975. *Practical Reasons and Norms*. Princeton, NJ: Princeton University Press.

—— 1979. The Obligation to Obey the Law. In Joseph Raz, *The Authority of Law: Essays on Law and Morality*. Oxford: Clarendon Press, 233–249.

—— 1984. Hart on Moral Rights and Legal Duties. *Oxford Journal of Legal Studies* 4: 123–131.

—— 1986. *The Morality of Freedom*. Oxford: Clarendon Press.

—— 1987. Government by Consent. In John W. Pennock and J. Roland Chapman (eds.), *Authority Revisited, Nomos* 29. New York: New York University Press.

—— 1989. Facing Up: A Reply. *Southern California Law Review* 62: 1153–1235.

—— 1990. Authority and Justification. In Joseph Raz (ed.), *Authority*. New York: New York University Press, 115–141.

—— 1994*a*. Liberating Duties. In Joseph Raz, *Ethics in the Public Domain: Essays in the Morality of Law and Politics*. Oxford: Clarendon Press, 29–43.

—— 1994*b*. Authority, Law and Morality. In Joseph Raz, *Ethics in the Public Domain: Essays in the Morality of Law and Politics*. Oxford: Clarendon Press, 210–237.

Routledge, Bryan R.. 1998. Economics of the Prisoner's Dilemma: A Background. In Danielson, Peter A. (ed.), *Modeling Rationality, Morality, and Evolution*. Oxford: Oxford University Press, 92–118.

Schelling, Thomas C. 1960. *The Strategy of Conflict*. Cambridge, MA: Harvard University Press.

Simmons, A. John. 1979. *Moral Principles and Political Obligations*. Princeton, NJ: Princeton University Press.

—— 1996*a*. Philosophical Anarchism. In John T. Sanders and Jan Narveson (eds.), *For and Against the State*. Lanham, MD: Rowman & Littlefield Publishers.

—— 1996*b*. Associative Political Obligations. *Ethics* 106: 247–273.

—— 2001*a*. The Principle of Fair Play. In John A. Simmons, *Justification and Legitimacy: Essays on Rights and Obligations*. Cambridge: Cambridge University Press, 1–26.

—— 2001*b*. Fair Play and Political Obligation: Twenty Years Later. In John A. Simmons, *Justification and Legitimacy: Essays on Rights and Obligations*. Cambridge: Cambridge University Press, 27–42.

Skyrms, Brian. 1990. *The Dynamics of Rational Deliberation*. Cambridge, MA: Harvard University Press.

—— 1996. *Evolution of the Social Contract*. Cambridge: Cambridge University Press.

—— 2004. *The Stag Hunt and the Evolution of Social Structure*. Cambridge: Cambridge University Press.

Sugden, Robert. 1986. *The Economics of Rights, Co-operation, and Welfare*. Oxford: Blackwell.

—— 2000. The Motivating Power of Expectations. In Julian Nida-Rumelin and Wolfgang Spohn (eds.), *Rationality, Rules, and Structure*. Dordrecht: Kluwer Academic Publishers, 103–129.

Ullmann-Margalit, Edna. 1977. *The Emergence of Norms*. Oxford: Clarendon Press.

Vanderschraaf, Peter. 1995. Convention as Correlated Equilibrium. *Erkenntnis* 42: 65–87.

—— 1998. Knowledge, Equilibrium and Convention. *Erkenntnis* 49: 337–369.

Verbeek, Bruno. 2008. Conventions and Moral Norms: The Legacy of Lewis. *Topoi* 27: 73–86.

Waldron, Jeremy. 1993. Special Ties and Natural Duties. *Philosophy and Public Affairs* 22: 3–30.

—— 1999. *Law and Disagreement*. Oxford: Clarendon Press.

—— 2003. Authority for Officials. In Lukas H. Meyer, Stanley L. Paulson, and Thomas W. Pogge (eds.), *Rights, Culture, and the Law: Themes from the Legal and Political Philosophy of Joseph Raz*. Oxford: Oxford University Press, 45–70.

Weber, Michael. 2003. The Reason to Contribute to Cooperative Schemes: An Examination of Christopher McMahon's "Collective Rationality and Collective Reasoning". *Philosophical Studies* 116: 171–181.

Winch, Peter. 1990. *The Idea of a Social Science and Its Relation to Philosophy* (2nd edition). London: Routledge.

Wolf, Robert Paul. 1970. *In Defense of Anarchism*. New York: Harper & Row Publishers.

Wolff, Jonathan. 1990–91. What Is the Problem of Political Obligation? *Proceedings of the Aristotelian Society* 91: 153–169.

—— 2000. Political Obligation: A Pluralistic Approach. In Maria Baghramian and Attracta Ingram (eds.), *Pluralism: The Philosophy and Politics of Diversity*. London: Routledge, 179–196.