Central European University

Political Science Department

A Reason-based Justification for Liberal-Democratic Authority

Ruzha Smilova

A Dissertation submitted to the Political Science Department of the Central European University in partial fulfillment of the requirements for the degree Philosophy Doctor in Political Science

Supervisor: Professor János Kis

Budapest, June 2005

Acknowledgements

My greatest thanks for this thesis go to my supervisor Professor János Kis, who has patiently provided me with support and inspiration through the years of my doctoral research. It would be truly embarrassing to indicate all the points in the thesis, which are my reaction (imperfect as it might be) to his invaluable ideas, comments, and critical remarks.

I would also like to express my gratitude to Professor Joseph Raz, who supervised my research at the University of Oxford in the academic 2000-2001. His friendly and encouraging guidance helped further clarify the positions, defended in my thesis. I specifically thank him for his generous help on Chapters Four, Five, and parts of Chapter Eight, as well as for the numerous discussions we had on his work and particularly on his account of authority.

The thesis would not have been possible without the excellent academic environment of the Political Science Department of the Central European University, at whose doctoral seminars I have presented my work. I would like to thank all the participants in these seminars. I specifically thank Loránd Ambrus-Lakatos, Gergely Bognár and Serhiy Pukas for their written comments and the stimulating discussion we had on what has become Chapter Six of my thesis. I also benefited significantly from discussions with Nenad Dimitrijevic and Ferenc Huoranszki, who in the initial stages of my work helped me focus my research. I am also grateful to Carol Harrington for going through the language of the final draft and for her helpful suggestions for its improvement.

My special thanks go to the Central European University for all the financial and logistical support, as well as to the United Kingdom's Foreign and Commonwealth Office for providing the Chevening Scholarship for my research at Oxford University.

Table of Contents

Acknowledgements i
Table of Contents ii

Introduction 1

Part One. Authority: Concept and Justification 7

Chapter One. Practical Authority and Protected Reasons for Action 7
1. Practical authority defined. 7
1. 1. Three Models of Authority Distinguished 9
1.2. Raz's Model of Practical Authority 10
2. Content-independent Reasons for Action 13
2.1. Defining Content-independent Reason for Action 13
2.1.1. "No Direct Connection" or "No Dependence on Evaluative Properties" requirement for CiR? Epistemic versus ontological interpretation of CiRs. 15
2.1.2. "No Direct Connection" Requirement 17
2.1.3. "No Dependence on Evaluative Properties" requirement 21
2.2. The Coherence of CiR: The "Normative Gap" Problem 22
3. Defining Exclusionary Reasons 24
3.1 The Coherence of the Concept of Second-order Reason for Action 25
3.1.1. Conformity v. Compliance with Reason 25
3.1.2. The Special Case of Agent-relative Deontological Reasons 27
3.2. The Coherence of ER as Negative *Second-order Reason* 28
3.2.1. The Partial Conflict Resolution Argument 30
3.3. Problems with Weight and Scope 33
4. Conclusion 40

Chapter Two. The Service Conception of Legitimate Authority: Normal Justification Thesis and the "Moral Duty to Obey" Problem 42
1. The Service Conception of Legitimate Authority 42
1.1.The Three Core Theses 42
2. The Normal Justification Thesis: Interpretations 46
2.1. Exclusively Substantive or Inclusive? The "Filtering" Role 46
2.2. Objective Only or a Subjective Element as Well? 50
2.3. Cumulative or One-shot Test of Legitimacy? 53
2.4. Maximising or Satisficing? 54
2.5. Turning "Oughts" into Duties? 56
2.6. Deference or Dialogic Model of Authority? 57
3. The Coherence of NJT: the Practical Difference Thesis 59
4. NJT and the "Moral Duty to Obey" Problem 63
4.1. The Goal-Independence Condition for Duty 64
4.2. An Instrumentally Justified Categorical Duty? 68
5. Conclusion 70

Part Two. The Case of Political Authority 72

Chapter Three. Normativity and Coercion 74
1. Introduction. The Disjunctive View of Normativity and Coercion 76
2. The Intuitive Argument: The Requests - Authoritative Utterances Analogy 76
3. The Substantive Arguments: Incompatibility of Threats and Authoritative Commands 81

3.1. Threats and Expressive Reasons*: "Expressive Significance of Deference" Argument* 82
3.1.1. Non-instrumental Reasons for Obedience: Deference in Order *to Express Respect* 82
3.1.2. "Shared Public Meaning of Deference" Argument 85
3.2. Reliance on Citizens' Good Will and Cooperation: "Importance of Being Trusted" Argument 86
3.2.1. A Formal Objection Rebutted 89
3.2.2. Substantive Objections 95
3.2.2.1. Fairness and Efficiency-based Coercion 95
3.2.2.2. The Assurance Rationale for Threats 99
3. 3. Preference for Compliance: "Gift Analogy" Argument 102
4. The Formal Argument: "Appeal to Inclinations" 104
5. Conclusion. The Disjunctive View: Unintended Consequences 109

Chapter Four. The Normative Supremacy Claim and the Autonomy Condition: A Critique of Ronald Dworkin's Endorsement Constraint Thesis 113
1. Endorsement Constraint Thesis Analysed 116
2. Challenge versus Impact Models of Critical Well-being: The Underlying Indexed versus Transcendent Value Distinction 123
3. Indexed Value: Weaker and Stronger Interpretations 127
3.1. The Two Interpretations 128
3.2. The Appeal of Indexed Value 130
4. Limitations and Parameters of Good Life 132
4.1. Normativity through Parameters? 134
4.2. Justice: The Universal Normative Parameter? 136
5. Cultural Paternalism and the Endorsement Constraint 139
6. Conclusion: Beyond Endorsement 145

Chapter Five. The Normative Supremacy Claim and the Autonomy Condition: A Defense of Agent-Relative Reasons for Action 148
1. Introduction: Autonomy and Agent-Relative Reasons for Action 148
2. Agent-Relative Reason for Action Defined: Structure and Main Types 151
3. Divergence of Value: Agent-Relative or Agent-Neutral Explanation? 154
3.1. Divergence of Value? 154
3.2. Divergence of Value via Agent-Relativity? 158
3.2.1.Divergence of Value via Presence of Goals 162
3.2.2. Goals Only or Desires as Well? 162
4. The Limits of the Argument. Some Objections Considered 166
5. A Reason-based Account of Autonomy 173
6. Conclusion: The Autonomy Condition and the Normative Supremacy Claim 175

Part Three. Authority and Instrumental Rationality 179

Chapter Six. The Rationality of Deciding to Follow Authority: The Toxin Puzzle Analogy 182
1. Toxin Puzzle and the Instrumental Justification of Authority 182
2. Analogy TP – Instrumental Rationality of Deciding to Follow Authority? 185
2.1. The Structure of TP 186
2. 2. Preliminary Objections 187
2.2.1. Two Levels of Decision 187
2.2.2. TP in Rationally Intending Plurality of Acts? 189
2.2.3. ER Solution to TP on Particular Occasion 190
2.2.4. Reappearance of the Puzzle? 191
3. "Autonomous Benefits" in Deciding to Follow Authority? 194
3.1. The Backward-Induction Argument 195

3.2. An Objection Considered 198
3.3. Limits of Knowledge and Rationality: Non-Autonomous Benefits 202
4. Autonomous Benefits in Deciding to Follow Authority 203
4.1. The Wrong Belief Case 203
4.2. The Ambiguous "Clear Mistake" Case 204
5. Conclusion 206

Chapter Seven. The Rationality of Deciding to Follow Authority: The Instability of the Instrumentally Justified Decision Strategy 208
1. Defining the Instability Problem 209
1.1. Clear/Great Mistake Distinction – Ambiguity in Clear Mistake Cases? 209
1.2. The "Instability" Problem 211
2. Is the Instrumentally Justified Decision-making Strategy Stable? 213
3. The Stable Instrumentally Justified Decision-making Strategy: Resolute or Modified Sophistication Choice? 216
3.1. The Resolute Strategy of Rational Choice 217
3.2. Bratman's Modified Sophistication Strategy - Apt Strategy for Raz's Account of Authority? 220
3.2.1 Bratman's Strategy 220
3.2.2. Applying Bratman's Strategy to Razian Authority? 222
4. Beyond the Decision Model: A Critique of Scott Shapiro's Constraint Model of Authority 226
5. Conclusion 232

Part Four: The Authority of a Liberal-Democratic Political Order 233

Chapter Eight. Content-independent Reasons for Action by Democratic Pedigree? 234
1. Introduction: Democratic Authority and the Content-independent Reasons Problem 234
2. CiRs: The Normative Gap Problem 237
2.1. The Normative Gap Problem and the Service Conception of Legitimacy 237
2.2. Closing the Gap: Appeal to Merit at a Next-order Level of Justification? 239
2.3. Raz's Autonomous Reasons Solution 240
3. Types of Strategies for Validity of CiRs Distinguished 243
3.1. Raz's Instrumentalist Outcome-based Strategy 245
3.1.1 Problems with Validity of ERs 245
3.1.2.Beyond Mere Rationality: Substantive Shortcomings of the Instrumentalist Strategy 248
3.2. The Proceduralist Strategy: Valid CiRs by Democratic Pedigree? 251
3.2.1. A Case of CiRs? 252
3.2.2. Are CiRs with Democratic Pedigree Valid? 254
3.2.2.1. Democratic Proceduralism 255
4. NJT's Filtering Role: The Full Legitimacy Test and Liberal Democracy 260
5. Conclusion 264

Conclusion 266

Bibliography 272

A Reason-based Justification for Liberal-Democratic Authority

Introduction

Can a reason-based justification for political authority in a liberal-democratic political order be offered? This was the question that prompted my journey into theories of authority. The focus of my present theoretical interests is the reason-based account of practical authority advanced by Joseph Raz. My initial interest in this particular conception was triggered by the fact that it offers the most sophisticated account of the concept of authority to date. My motivation to continue analysing it in detail was strengthened by the fact that it offers, as a test of legitimacy, not only a reason-based account of the concept of practical authority in general, but specifically a reason-based type of justification for political authority as well. This perfectly fitted my initial search for a reason-based type of justification for political authority.

Joseph Raz's conception of legitimate authority grounds the justification of political authority in the sound reasons of its subjects. When authority overall brings improved conformity to those reasons, its exercise is justified: it is a legitimate authority. Thus this is a reason-based, and not will-based type of justification, since whether the exercise of authority is or is not justified does not depend on whether its subjects agree with its orders, or agree that the above test of legitimacy is met. The justification of political authority is a matter of objective reasons, and it is a matter of objectively improved conformity to those objective reasons. As stated (and it is understandably simplified), this position seems unnecessarily rigid, but it always helps to be as explicit as possible in the beginning of a discussion, as to what is ultimately at stake. The interesting question is whether a reason-based type of justification (of the above-mentioned type, in a more or a less rigid form) can account for the authority of a liberal-democratic political order. However, my main question in the present thesis is thus not the general one: can a reason-based justification be offered for a liberal-democratic authority. Rather, I focus on Joseph Raz's account of authority - the most sophisticated and fully developed, among the reason-based ones, and ask whether it has the resources to provide such a justification. This issue is addressed in the course of a long, often circuitous discussion of the main building blocks of his conception. I believe, nevertheless, that this is necessary

for providing as rich a picture of this conception, with its main theoretical advantages as well as attending problems, as possible. I first define the fundamental concepts, some of them unique to this conception of authority, then identify the problems with the most controversial of them. I next evaluate the success of the critiques levelled against them, and offer my own arguments more often as a critique, but sometimes as a defence of some of the main tenets of this conception.

My answer to the question: does Raz's account of authority provide a fully adequate account of the legitimacy of a liberal-democratic type of authority, is negative. There are problems with this account of legitimacy already at the general level. The exclusively instrumentalist in character legitimacy test it advances is not perfectly congruent with the common-sense understanding of legitimate authority correlated with a duty to obey. Nor does it deliver on its promise to provide an unambiguous solution to the latent in our concept of authority rationality paradox: namely, authority is either wrong or superfluous, so obeying it is never rational. There are also specific problems with it as an account of the legitimacy of political authorities. I do recognise, however, the need to accommodate its sound points within a non-instrumentalist reason-based account of the justification for a liberal-democratic type of authority.

In my thesis, I proceed as follows. In the first chapter, I introduce Joseph Raz's account of authority by contrasting it with alternative models of authority. After outlining the advantages of construing authority on the model of practical authority this author meticulously develops, I offer an analysis of its main concepts: that of protected reason for action with its two components – content-independent and exclusionary reasons for action. For Raz, practical authority claims to create valid protected reasons for action for its subjects: they indeed are valid whenever authority is legitimate. Before addressing the justification question – when, under what conditions are those reasons valid, and authority legitimate, I examine the theoretical difficulties with establishing the coherence of the two components of the protected reasons concept. Thus, an analysis of the elusive concept of content-independent reason for action is offered. The main, "normative gap" problem with it: that what one ought to do when one has such a reason for action does not depend on the *good* of acting as required by it - the validity of such a reason does not depend on the *evaluative* properties of the action it requires, is then identified, and briefly

discussed. Next follows the even more contested concept of an exclusionary reason for action. The doubts concerning its status as a second-order, and negative reason for action are addressed seriatim. Further, some not yet sufficiently analysed problems with determining the weight and the scope of such reasons for action are identified and their implications discussed.

Despite the fact that there are problems already at this conceptual level, I move to the level of justification, where my ultimate interest lies. Thus the task of my second chapter is to introduce Joseph Raz's Service conception of legitimate authority, as well as to outline what I believe to be the main theoretical problems with it. This conception conditions the legitimacy of authority on it bringing improved conformity to subjects' own reasons. Thus Raz's Service conception of authority is not only reason-based, but has a generally instrumentalist character, employing a maximising account of instrumental rationality. This conception consists of two moral (the Normal Justification and the Dependence), and a structural (the Preemption) theses, and is subject to meeting the requirements of the autonomy condition. I discuss these elements in the context of its main ambition. It is to dissolve the paradox of practical rationality that plagues the traditional common-sense concept of practical authority. The focus of the chapter is on the instrumentalist strategy – the Normal Justification Thesis, for dissolving this paradox. After outlining the main interpretations of this thesis, I raise some concerns about its coherence. I then identify the main problem with it: it cannot account for the sense in which practical authority, when legitimate, makes practical difference to what its subjects ought to do by giving rise to a moral duty for them to obey its orders. On unrestricted instrumentalist grounds it might be possible to show that authority gives a new hypothetical rational requirement to obey, but it is a further, and serious problem for this theory to show how authority can turn this rational requirement into a *moral duty* to obey.

In the second part of my thesis, I already focus on the specific features of *political authority*, the central case of the more general practical authority. Some of the essential features of political authority, and of its main instrument – the law, are, according to Raz, that it necessarily makes a normative claim to authoritativeness, i.e. claims to be a legitimate authority, and that it necessarily claims comprehensive supremacy over all other normative domains - with no limits other than the limits it itself recognises. A

further central feature of law and the state, though not recognised as their *essential* feature by Raz, is law's and the state's undeniable and extensive use of coercion. My concern in this part is to explore the mutual compatibility of these central features of political authority, as well as to see whether and how they fit within Raz's general conception of practical authority.

In the first chapter of this part I discuss in detail the case Meir Dan-Cohen makes for his disjunctive view of normativity and coercion. The question that drives my analysis is whether his arguments were successfully responded to, or rather, he brought out serious, not yet fully appreciated and tackled with problems for the compatibility of the normative claim to legitimacy political authority necessarily makes, on Raz's account, with state's extensive use of coercion. The remaining two chapters of this part concentrate on the compatibility of two other central elements of Raz's account. These are the normative supremacy claim the state (through its main mechanism – the law) necessarily makes, and the autonomy condition of the Service conception, which states that the legitimacy of authority is conditional on showing that deciding correctly is more important than deciding for oneself. I discuss the plausibility of two theses - the endorsement constraint thesis and the agent-relative reasons thesis - as justifications for the autonomy condition. While I show that there are serious problems with the first thesis, as famously defended by Ronald Dworkin, in the last chapter from this part I offer a defense of the second, agent-relative reasons justification for the autonomy condition. The question, which drives the whole discussion of the limits the autonomy condition (in its more plausible interpretation) imposes on the legitimate exercise of political authority, is addressed at the end of this part. I ask there whether Raz's analysis of the concept of legal and political authority can account for what is specific about political authority in a liberal-democratic political order, on which the autonomy condition is believed rightly to impose *external* limits on political authority, irrespective of whether this is recognized by the latter or not. More particularly, I ask whether the claim to supremacy overall all other normative domains, an essential feature of legal and political authority according to this analysis, is characteristic of this special type of authority. Finally, I ask whether the internal coherence of Raz's conception is not also compromised. The general tenor of the Service conception of practical authority – obedience to authority is ultimately justified

only when licensed by practical reason/morality, seems to go against this normative supremacy claim as a central feature of political and legal authority. Discussing this major issue constitutes the concluding section of this part of my thesis.

The third part looks more closely at the already identified as problematic aspect of the Service conception: its instrumental character, with its inherent maximising logic. This characteristic holds the promise of solving the rationality paradox of obeying authority, and it is here that I ask whether this conception delivers on its promise and indeed does have such an advantage. The more specific question I address is whether it is individually rational to decide to follow an instrumentally justified authority, if to follow authority means to take its directives as protected reasons for action. In the first chapter of this part I evaluate a suggestion that *deciding* to follow authority might not be rational in the same way as deciding (and forming an intention) to drink the toxin in Gregory Kavka's famous Toxin Puzzle is not rational. Underlying this possible analogy is the fact that the maximising logic of rationality seems not to permit acting in sub-optimal ways. The analysis in this chapter helps illuminate another problem for the rationality of deciding to follow authority, which I discuss in detail in the second chapter. Does the maximising logic, inherent in the instrumental conception of rationality used by Raz's conception of authority, undermine rather than uphold its capacity to solve the rationality paradox? The more specific question I ask at this point is: is the strategy of always following what one believes to be a legitimate authority overall, even when one disagrees with its directives on a particular occasion and happens to be right to disagree (since the authority did not get the balance of ex ante reasons right) rational? And if rationality requires allowing room for exceptions, does such a rational strategy have resources to solve the "instability problem": if one is always tempted in cases of disagreement with authority to disregard its directives, and gives in often enough to this temptation, one ends up being worse off by deciding to follow authority than if one always followed one's own judgment only instead? Does not that suggest that deciding to follow authority is not rational if no stable decision-making strategy is available? I closely analyse Raz's own account, as well as ask whether the rational strategy developed by Michael Bratman to solve similar rationality problems of dynamic choice, could provide, when applied to this account, a plausible solution to the instability problem. At the end of this part I discuss the success

of an alternative model of authority, developed by Scott Shapiro in response to the same problems. This model breaks radically free from the main presuppositions of the traditional models of authority, responsible for their inherent rationality puzzles. The question I ask is whether the solution this model offers is indeed successful.

The concerns raised in the third part are added to the previous, more general critique against Raz's instrumentalist justification of political authority - that this type of justification has problems accounting for a central notion of authority: that one has a *moral* duty to obey legitimate political authority, acting within the bounds of its jurisdiction. The rationality problems, together with the plausibility of this latter critique warrant, I believe, trying to develop an alternative type of justification in the case of political authority.

Thus the question I ask in the fourth, concluding part of my thesis is: can a plausible case be made for a liberal-democratic form of political authority, on the ground that the protected reasons for action authority necessarily claims to give to its subjects, can be valid when given by such authority? I here explore the potential of democratic authority to provide valid content-independent reasons for action (the first component of the protected reasons for action concept): they have been challenged precisely on the ground that they cannot be valid, and acting on them – rational.

I claim that there can indeed be valid *content-independent reasons* by democratic pedigree. I try to identify the explanation for the success of democratic authority in this regard. I ask whether it has to do with abandoning the accounts of the justification of authority, premising legitimacy on authority being instrumental for achieving maximally improved conformity to practical reason. The failure within this theoretical framework to explain the validity of content-independent reasons may confirm this hypothesis.

And finally, I ask whether a plausible account of the legitimacy of liberal-democratic authority can accommodate the sound points of Raz's Service conception – that following authority is often justified on instrumental grounds. In light of the above discussion of its central elements, and their compatibility with Raz's account of practical authority, the Normal Justification thesis may need to be modified, and downgraded into being just a necessary condition for a plausible justification of authority. The path to be taken towards its modification is roughly outlined in the concluding sections of my thesis.

Part One
Authority: Concept and Justification

Chapter One
Practical Authority and Protected Reasons for Action

1. Practical Authority Defined.

To have authority is to have a right to rule – to have a right to impose one's will on one's subjects. How could authority's claim to have a right to impose its will on its subjects ever be justified? It is an established tradition to start a work on the legitimacy of political authority by discussing the position of the philosophical anarchist, implicit in this rhetorical question. The challenge Robert Paul Wolff in his *In Defense of Anarchism* and the philosophical anarchists following his lead pressed, is that the notion of legitimate, or *de jure,* authority is incoherent. The concept of legitimate authority used is that of practical authority, or authority over actions, not beliefs. This is the concept of authority accepted both by philosophical anarchists and some of their critics of otherwise widely diverging persuasions[1]: it is believed to nicely capture what is distinctive of authority. Political authority is, accordingly, just a species of this more general type of authority. The distinctive feature of authority of this practical type is that it has the normative power to guide the behaviour of its subjects by affecting their practical reasoning as to what they ought to do. In this, authority is distinct from mere power[2]: the ability to change what one's subjects do, by compelling conformity to authority through the use of physical force, manipulation, or some other non-normative means. Authority, unlike the proverbial gunman, does not simply change what its subjects do, but it changes what they *ought to* do: it changes their normative situation by changing the reasons that apply to them.

---

[1] Together with philosophical anarchists such as Wolff and Simmons, the concept of political/legal authority as a species of the more general type of *practical* authority is accepted by legal philosophers with hard (exclusive) positivist - Raz, Shapiro (2002a), Marmor (2001) as well as soft (inclusive) positivist - Coleman (2001), Himma (2000), Waluchow (2000), and even natural law - Finnis (1986) views on the nature of law. For a view opposing the practical authority model of political and legal authority, see Hurd (1999), Regan (1989), Moore (1989) Alexander (1982), among others.

[2] Wolff (1970).

To have practical authority is to have a right to rule, implying a general obligation on the part of its subjects to obey. The obligation to obey is usually construed as "prima facie, comprehensively applicable, universally borne, and content-independent."[3] Thus authority demands general obedience from each of its subjects on all occasions in a special, content-independent way: the action it commands has to be done because of the command and not for any other reason.

This description of authority with its demands for obedience puts authority in direct conflict with autonomous agents, who are, according to Wolff, required always themselves to determine what they ought to do, and never do what told simply for the reason that they have been told so. The contention is that a person, in obeying the commands of authority, cannot remain autonomous: even when he is acting correctly on the balance of reasons, he does it not because of the merit of so acting, but because he has been so commanded. This challenge to authority by the philosophical anarchist has been dubbed the "autonomy paradox," and has been distinguished from the "rationality paradox", sometimes taken as a more general instance of the same paradox.[4] The rationality paradox exhibits the alleged incompatibility of authority and rationality: authority, contrary to its claims, never gives valid reasons for action to its subjects. This is so not only when authority's directives are wrong (then acting on them is not rationally and/or morally justified), but even when they are right - because it is right reason/morality and not authority that directs subjects' actions in this latter case. It is because authority is either rationally/morally unjustified or redundant, that legitimate authority is once again an incoherent notion.

The influential account of authority advanced by Joseph Raz more than thirty years ago and meticulously developed and defended during the years, offers an analysis of the structure of authority (understood along the above lines), its role in the practical reasoning of the subjects and the conditions under which such authority is legitimate. It is a response to the anarchist challenge: it addresses both paradoxes above, though it takes the rationality paradox to be the main concern for developing a coherent account of practical authority. Thus, the concept of legitimate authority is, according to Raz's

[3] Kramer (2005: 179) quoted in Edmundson (2004: 215).

[4] In distinguishing two paradoxes of authority rather than one, as well as in their concrete characterisation, I follow Shapiro (2002a: 385-393). This author argues that the two paradoxes are irreducibly distinct.

analysis, not incoherent: authority can in principle be legitimate, it can provide its subjects with valid reasons for action of a distinct character. It could do this, if certain (presumably not impossible) conditions are met. I start my discussion of the success of Raz's account of authority by contrasting it with alternative accounts of authority, portraying it either as practical, though in a different sense than in Raz's account, or as theoretical only. In the process, the main concepts of Raz's analysis, to be extensively used and discussed later in my dissertation, are introduced.

1.1.Three Models of Authority Distinguished

The alternative models of authority offer different interpretations of the authoritative utterances and the types of reasons those utterances are taken to give to the subjects of the authority.[5] The model of *theoretical* (or recognitional[6]) authority, for example, takes the authoritative utterances ("X ought to F") as an *advise*, meant to provide X with a *reason to believe* that there are pre-existing (i.e. prior to the authority's directive) reasons that X ought to F. Such authority does not and is not meant to change the balance of reasons for or against F-ing – rather, it only indicates what those reasons with their right balance, are. This account of authority implies *the no difference thesis:* authority does not make a difference to how its subjects ought to act. There are considerable difficulties with this model, the main being that it does not account for the common sense notion of authority as itself imposing obligations on its subjects rather than simply informing them about their pre-existing obligations. Imposing obligations as a minimum requires that authority does provide its subjects with *new* reasons for action, rather than reasons for belief.

The model of *influential* authority interprets the authoritative utterance ("X ought to F") as a *request*, which request does already provide X with a reason to F. X's reason to F is a content-independent one: it is a reason to F because the authority has requested so, and not because the balance of reasons independently of authority's utterance directs one to do so. The new reason authority provides here is to be added to the pre-existent balance of reasons, thus potentially at least "influencing" it. Other things being equal, the new

[5] Raz's analysis of the models of authority: practical, theoretical and influential is in Raz (1979: 13-15), (1979: 21-22), and (1986: chapter 2). Hurd (1999) offers a detailed discussion and a defence of a version of the theoretical model as the only coherent one among others. She provides a very useful systematisation and characterisation of the three types of authority, which I follow in my text.

[6] Raz (1986: 28).

reason authority provides determines how subjects ought to act: it serves as a tie-breaker, and thus *makes a practical difference* to how its subjects should act. It can, moreover, if it is strong enough, outweigh even strong reasons against the commanded by the authority action. Because of all this, authority on this "influence" model is a species of practical authority widely construed. Nevertheless, this model falls short of accounting for the sense in which authority has a right to impose *obligations* on its subjects, and not simply provide them with new *reasons* for action. This is what the model of practical authority takes as its task to remedy. On it the authority not only gives new content-independent reasons for action, but makes these binding for the subjects by "protecting" them with a further, special type of reason for action.

1.2. Raz's Model of Practical Authority

Thus the ambition of the *model of practical authority* narrowly construed is precisely this: to account for the sense in which authority, when legitimate, has a right to impose obligations on its subjects. On it, the authoritative utterances are already interpreted as *orders* and not simply as requests. It takes authoritative directives to be obligatory for its subjects – they provide subjects with *protected* reasons for action. This is the account of authority advanced by Joseph Raz.[7] The protected reasons for action provided by authoritative utterances (as well as by decisions, mandatory rules, norms, and promises[8]) comprise both a first-order content-independent reason for the action required by the directive, and a second-order exclusionary reason, excluding some reasons against the required action. [9]

The difference with the model of "influential" authority is clear: the latter provides subjects only with new first-order, content-independent reasons for action, which are to be *added* to the pre-existing content-dependent reasons. It does not give them exclusionary reasons. These reasons are meant to "protect" the newly added by the authority reasons by excluding acting on the reasons that contradict the command. It is precisely this "protecting" role, played by the exclusionary reasons, which can explain how the authoritative reasons can create obligations, and not simply add first-order

[7] In Raz (1979: chapter 1), and Raz (1986: chapters 2-4).
[8] Raz (1989: 1160).
[9] For this definition of exclusionary reason, see Raz (1990b: 39).

reasons for the subjects to obey. This difference (*excluding*, not simply *adding* reasons, is what distinguishes practical authority from other modes of affecting the practical reasoning of subjects) has considerable implications for the structure of the practical reasoning of the subjects. On the "practical" model in the narrow sense, the decision the subjects make about what they ought to do, when an authoritative utterance is addressed to them, is a decision "all things considered." Thus both the first-order non-excluded reasons and the protected reason with both its elements – the first-order content-independent reason given by the authoritative utterance and the second-order exclusionary reason, determine "what ought to be done, all things considered." On the "influential" model, in contrast, this decision is "on the balance of first-order reasons alone" (the pre-existent content-dependent reasons plus the newly added by the authority content-independent reason).

The advantage of the model of practical authority narrowly understood is that by taking commands to provide subjects with protected reasons for action it can better account for the *binding*, obligatory force of the authoritative directives, issued by political authorities. It accounts for the sense in which authority, when legitimate, has indeed a right to impose obligations on its subjects. Were one only to *add* the reasons provided by the authoritative utterance (which is the case when the authority *requests* that X do F) to the other pre-existing first-order reasons, and weigh them against each other, without excluding the underlying the authoritative directives reasons, the reasons provided by the authority would not have the property of categoricity, or obligatoriness, commonly attributed to them. Thus the claim of the authority that its directives be taken as binding, would be from the very beginning blunted. If the reasons, given by the authoritative utterances, are to be always weighed against the pre-existing reasons (so that the former can only consequently, as a result of this weighing, be taken as binding), it cannot plausibly be claimed that the utterances giving those reasons are in any strong sense authoritative. For any utterances to be considered authoritative, the reason it provides should necessarily be regarded as categorical one: imposing an obligation, and not simply as a content-independent reason (no matter how strong the latter is presumed to be). This is at least the position of Joseph Raz, for whom the best conception of the authoritative,

or categorical character of the reasons provided by an authority holds that authoritative reasons are protected reasons, i.e. containing not only a content-independent, but an exclusionary reason as well. That the authoritative reasons are protected means that they win out over the pre-existing reasons not by weight (in which case the new content-independent reasons would have only a presumptive and not exclusionary force), but by kind. Raz's conclusion is that the correct analysis of authority, which brings out its central features and exhibits its internal structure, as well as its role in the practical reasoning of its subjects, describes it as practical authority in the narrow sense specified above.

The Razian model of practical authority has been subjected to strong critiques on the ground that either there are no valid exclusionary (and by implication, no protected reasons as well) reasons, or that the very concept of exclusionary reason is incoherent.

Those critics, who agree that distinctive of authority is that their directives provide content-independent reasons for action, opt for the model of influential authority.[10] More radically, some critics challenge even the content-independent character of the authoritative reasons. They deny that authoritative utterances provide *new* reasons for action, thus denying that they provide any *reasons for action* at all. They defend the model of theoretical authority, where the claim that authority gives new reasons for action is substituted with the position that authority gives *reasons for belief* in the validity of some pre-existing reasons for action.

The task of defending the model of practical authority is to demonstrate how authority can ever provide its subjects with *valid* protected reasons for action. Before that, however, it must be shown that the very concept of protected reason for action is *not incoherent.*

Thus there are two main challenges to the model of practical authority, deriving from the difficulties with demonstrating that there are valid protected reasons for action. The first identifies the problem as the supposed impossibility of showing that there can in principle be valid *exclusionary* reasons for action. The second digs deeper, questioning whether

[10] This is suggested by Moore (1989). Frederick Schauer's position about the authoritative reasons (rules) having only presumptive force, i.e. sometimes but not always excluding considering all the reasons, which ground the exclusion, may be taken as an intermediary position between the models of practical (narrowly understood) and "influential" authority. See Schauer (1993: 88 – 93).

there can in principle be even valid *content-independent* reasons for action. The second critique affects the model of "influential" authority as well, arguably leaving as the only viable account of authority the model of theoretical authority. An evaluation of the plausibility of the model of practical authority in its own terms and as a response to the philosophical anarchist's challenge warrants a detailed discussion of these critiques. Carefully defining the main concepts used is a necessary prerequisite for this task.

Thus the protected reason for action authority claims to provide through its directive comprises a content-independent reason for action and an exclusionary reason against acting on (some of the) inimical to the commanded action reasons. Let us focus on the first: what is precisely a content-independent reason for action (henceforth CiR)?

2. Content-Independent Reasons for Action

2.1. Defining Content-independent Reason for Action

Raz's definition of CiR is:

> (A) "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."[11]

There is a growing literature on the problems surrounding the concept of CiR.[12] The talk of content-independence of legal norms, for example, is misleading, according to John Gardner,[13] since it concentrates (1) on the content of the norm, forgetting its form and (2) also disregards what is important there – that the dependence on the merit of the

[11] Raz (1986: 35)

[12] The concept CiR was introduced by H.L.A. Hart (1982). John Gardner (2001) takes Hart's definition to be misleading and not supported by Hart's own discussion of the same issues in Hart (1961). Gardner believes that nothing in Hart's work commits him to denying that the validity of legal norms provided by authority/law "can depend on their content so long as it does not depend on the *merits* of their content" Gardner (2001: 213, emphasis in the text). Gardner finds support in Hart's work to go even further: to the extent that the reasons provided by an authority/law are *not dependent on both the merit of their content and their form*, they will have their distinct character. Markwick (2000: 579 – 596) attacks both the coherence and the distinctness of CiRs. A detailed analysis of this concept he offers in Markwick (2003), where he challenges the accepted view of many legal theorists that content-independent reasons (exemplified typically but not exclusively by legal reasons) do exist.

[13] Gardner (2001: 213)

content,[14] and not simply dependence on content is to be avoided, if authoritative reasons are to be distinct.

Nothing in my discussion below will be gained (nor will be substantially changed) by substituting CiR with the suggested by Gardner term "merit-independence" of authoritative reasons. Since "content-independence" as distinctive of authoritative reasons has gained currency in the debates, I will stick to the accepted usage, taking into account the sound points of Gardner's critique. So, an alternative definition of CiR, after Gardner, would be:

> **CiR** is a reason, in which there is no dependence on the merit of the content of the action, for which it is a reason.

Since I find the expression "no dependence on the merit of the content of the action" cumbersome, let me substitute it with, what seems to me, a more neat, though no less clear expression: "no dependence on the evaluative properties of the action." Thus we get

> (B) CiR is a reason for action, which is not dependent on the evaluative properties of the content of the action, for which it is a reason.

Before going into the details of these alternative definitions, let me note that Raz's definition has one clear advantage. It is much richer in not simply explaining the meaning of the concept by way of negation (saying what it is not), but by also providing some positive account of what it involves. Firstly, the "say so" of a person is what produces this peculiar reason. Secondly, its validity is no less peculiar in being immune to the counterfactual variance of that person's "say so." A plausible and full definition of CiR should preserve these two positive characterisations. The definitional problems I discuss below concern only the first part: whether CiR is negatively characterised by "no direct connection" with the required action (A), or with "no dependence" on the evaluative properties of its content (B).

---

[14] Gardner talks of not relying on the merits of the form (and not only of the content) of the legal norm. This talk of merit of the form is at home within the framework of legal norms that Gardner is discussing in the quoted text. Legal norms have both distinct form (e.g. compliance with rule-of-law standards, generality, prospectivity, etc.) and content, and can have/lack merits regarding both of them. In other context I find it hard to understand how the form of an action can in principle have/lack merits, which is distinct from the merits of its content. It seems always possible to re-describe the content of an action so that it will include everything, allegedly pertaining to its form. Thus, I choose in my text to talk about evaluative properties of action, and leave aside the problems of distinguishing evaluative properties pertaining to the form and those pertaining to the content of actions.

2.1.1. "No Direct Connection" or "No Dependence on Evaluative Properties" requirement for CiR? Epistemic versus ontological interpretation of CiR.

The two requirements (A) and (B) in the above definitions are *not identical*. The requirement (A) that there be "*no direct* connection between the reason and the action for which it is a reason", allows for distinguishing it from a more indiscriminate requirement that there be *no such connection whatsoever*. Compare this with the requirement (B) that there be "*no dependence* on the evaluative properties of the action". (A) may allow for some indirect connection between the reason for action and some *merit* of acting on this reason, which merit does not reside in the evaluative properties of the required action itself), while (B) may not. This shows that the two do *not necessarily have the same implications*. Further, one might think the qualification that there only be *no "direct" connection* between the merit of the action and the reason to perform it, very important for understanding both how there can be valid CiRs and for how the concept of CiR is not incoherent. I show that there are difficulties with this suggestion. So, a more careful analysis of the meaning of the concept of CiR, of the "no direct connection" requirement, and its relation to the "no dependence on the evaluative properties of the action" requirement in particular seems necessary,[15] in order to account for what is distinct about CiRs.

As a first step in this regard, let me offer two distinct interpretations of the claim that authoritative commands give CiRs to its subjects: it will help unpack the meaning of this concept, and determine its requirements. The first, epistemic reading is that the *identification* of authoritative reasons does not involve appeal to the merit (evaluative properties) of the requested action itself: no evaluative, content-dependent argument is involved in the *identification* of these reasons. This reading could support an indirect argument for understanding authoritative directives as CiRs. This argument may run as follows: were the identification of authoritative reasons to involve appeal to the merit of the requested action, there would be no difference between ordinary content-dependent

[15] Some problems with understanding the precise meaning of the "no direct connection" requirement for content-independent reasons are discussed by Markwick (2000: 579 – 596). For a suggestion that there is a link of the "no direct connection" requirement with the Service conception of authority, and the Pre-emption thesis in particular, with its requirement that authoritative utterances (the arbitrator's decisions) be identifiable independently of their underlying reasons, see Shiner (1992: 52-53).

reasons and reasons, provided by an authoritative utterance. Phenomenologically we distinguish between ordinary reasons and "authoritative" reasons, precisely on the ground that we do not think necessary to inquire into and "calculate" the merits of the action, required by an authoritative utterance, in order to decide whether we ought to perform it. The distinction content-dependence/content-independence of reasons is one way of accounting for this phenomenologically observed difference. So, there is a prima facie case for the existence of CiRs as distinct from other reasons. This indirect "argument from phenomenology" is clearly not in itself sufficient to demonstrate that CiRs do exist: there could be a different explanation than in terms of CiRs for our experience when faced with authoritative directives.

The second, I believe more interesting, reading of the claim that authoritative utterances give CiRs is the *ontological* claim that these authoritative directives do indeed provide valid CiRs: CiRs as distinct from content-dependent reasons *do exist*. Raz believes this can be established by providing a compelling rationale for their existence.

For this purpose he advances a second, "teleological argument."[16] Raz's explanation of how taking commands as CiR (considering the merits of the action is not necessary in deciding to perform it) can be rationally justified is a main support for the observed difference between the two types of reasons. This difference plays a crucial role in the explanation. It is only through acting on the reasons provided by authority without balancing their merits (taking them as CiR) that one could meet the raison d'etre of authority: which, we will see, consists in bringing improved conformity to one's own reasons. If acting on CiR required appeal to the merit of the action, this advantage would be lost.[17]

The CiR thesis is not simply a pragmatic ("it is useful" to take authoritative commands as CiR), nor merely an epistemic (the way we "realize" the existence of certain reasons is by not considering the merits of the actions, for which they are reasons) thesis. It is an

[16] For discussion of Raz's phenomenological and teleological arguments, see Edmundson (1993: 339-342). The label "teleological" was introduced by Edmundson (1993: 340). I think it nicely captures the gist of this argument.

[17] Let me just note here in passing that the two arguments: the phenomenological and the teleological, are used by Raz in the defense of the exclusionary reasons for action as well. The reason why the arguments for the CiR and exclusionary reasons are parallel, is that Raz explains the plausibility of the concept of CiR, and indicates how such reasons can be valid, by introducing the concept of exclusionary reason.

ontological one: the validity of authoritative reasons does not depend on the merits (or the nature more broadly) of the action commanded, but on the fact that they are issued by an authority. This explains how authority can make a practical difference to how its subjects should act. Further, it is the ontological thesis, which upholds the pragmatic and the epistemic ones. It is because the authoritative reason's validity does not depend on the merits of the action commanded, that one does not need to inquire into these merits in order to decide whether one actually has this reason (epistemic thesis), nor does one stand to gain anything by doing so (pragmatic thesis).

Distinguishing these two interpretations of the claim that authoritative reasons are CiRs (the epistemic - *identification*-without-recourse-to-merit and the ontological - *validity*-without-dependence-on-merit) helps us better understand CiRs. Not only are the two interpretations distinct, but the former can be more easily associated with the "no direct connection" requirement in definition A above, while the latter seems closer to the "no dependence on the evaluative properties/merit" one (in B). The primary in importance ontological thesis, explaining how authoritative reasons can make practical difference to subjects' reasons for action, better fits requirement (B) and the second interpretation of having "validity without dependence on merit."

Are there further grounds for preferring (B) over (A) requirement? The requirement (A) that there be "*no direct* connection between the reason and the action for which it is a reason", allows for distinguishing it from a more indiscriminate requirement that there be *no such connection whatsoever*. The (B) requirement is stronger - "*no dependence* on the evaluative properties of the action" may not even allow appealing to any *merit* of acting on the reason, even if that merit does not reside in the evaluative properties of the required action itself.

2.1.2. "No Direct Connection" Requirement

The "no direct connection" requirement (A) has an initial advantage: that there only be no "direct" connection between the merit of the action and the reason to perform it, may be thought important for understanding both how there can be valid CiRs and for how the concept of CiR itself is not incoherent. I show that there are difficulties with this suggestion.

The coherence of the concept of CiR is threatened unless it allows at least for some "indirect" connection of the reason to some merit/value. The concept of reason for action implies that there is some merit (value, good) in performing the action for which it is a reason. If this merit is not in the prospective value of the directly required action, there should nevertheless be some explanation as to why it would be desirable/good to perform this action. That there should be some merit in performing the requested action (which may or may not reside in the merit of the requested act-token itself), is reinforced by the claim that CiRs can in principle have strong weight and/or be exclusionary (as in the case of the commands of practical authority) reasons. If this claim is to be plausible, then, even if the merit of such reasons is to be independent from the content of the actually requested act, such reasons should be (indirectly) connected to some other merit, and *sizeable merit* at that. Thus showing that there can be merit in acting on CiR is important for showing that the concept of CiR is not incoherent.

All the above considerations explain the initial plausibility of requirement A. It trades on the possibility of distinguishing between an act-type and action-token.[18] It denies CiR's dependence on the merit of the actually commanded action-token ("X is legally required to pay 45 % income tax"). It does not deny its dependence on the merit, if any, of the act-type ("each citizen is legally required to pay income tax"), or its dependence on yet some further merit ("paying taxes is socially beneficial"). The merit of the act-type could be sizeable, thus arguably redeeming the coherence of the CiR concept.

This sizeable merit could potentially come from two sources. The first was already pointed at above: the merit of acting on a particular occasion as required may stem from the overall merit of performing the act type, of which the particular act is an instance.

(Ex) The merit of X doing f (instance of a class F) because Y required it, may lie in the merit of X doing F (i.e. whatever Y required). [19]

[18] A possibility of defining CiR, using the distinction act-type/tokens of the act-type is suggested by Markwick: "A legal reason to perform a certain act-type is content-independent since there would be reason to perform a particular token of this act-type even if this token had different properties" Markwick (2000: 594).

[19] A problem with this suggestion is that it assumes that we have CiR at the level of X's reason to f (act-token), and not at the level of X's reason to F (act-type). According to Markwick's definition above (see the preceding note), it is the reason for the act-type that is content-independent (and by implication, so is the reason for the act-token as well). If this is so, we will only have one possible source of merit for the commanded action: its pedigree.

The second, alternatively, refers not to the merit of the act-type, but to the merit of authority itself, which is the ultimate "source" of the reasons in question. The merit, in short, is in a *right* pedigree, stemming from a *good, meritorious* source.

The reasoning in this second case runs as follows. The authoritative reasons are provided by commands issued by an authority. The fact that something is commanded does not in itself constitute merit. For a command to be constitutive of the merit of performing the commanded action, the command itself should be issued by an authority, which is justified in issuing commands (thus obviously having some merits), and does not simply claim to be such. The very fact, that something claims to be a legitimate authority,[20] is not in itself merit. If this is so, then the merit of CiRs, explaining their potentially great weight/exclusionary force, cannot be explained by the fact that they have been issued by any authority, but by the fact that they have been issued by an authority which has some (sizeable) merit. An authority that has some sizeable merit is a more or less legitimate authority: it serves its subjects by acting within its own jurisdiction.

Whether these are two distinct sources of merit rather than one (is not one simply derivative from the other?), and what is the relation between them, is a complex issue. It is not important for evaluating the success of this particular response to the critique of the coherence of the CiR concept. Both sources of merit, irrespective of their interrelations, will satisfy the "no direct connection" requirement: since on both the connection with merit is *not established at the level of the concrete action required*, but is found at the second-order (merit of act type), third-order (merit of authority), etc. level, this connection is not direct.

The "no direct connection" requirement is, however, suspect. My contention is that this requirement only allows to account for the coherence of the concept of CiRs, at the expense of CiRs distinctness.

[20] An authority can sincerely claim to be a legitimate authority only if it can effectively issue commands – i.e. only if it is habitually obeyed and taken at least as a de facto authority by its subjects. In this sense even an authority that is not legitimate has the merit of being a de facto authority: an authority that is habitually obeyed. One should resist this conclusion, however. While it is true that there cannot be a political authority that is legitimate without being a de facto authority: being a de facto authority is an enabling, may be even necessary condition for legitimacy (in the political domain), it is also true that de facto authority is no more than that. The "merit" of being a de facto authority is entirely parasitic on the purposes the authority serves, and is never in itself sufficient to establish authority's merit.

To see this, consider first the following argument against the coherence of CiR. If one derives the sought-for merit of following authoritative commands from authority being justified, i.e. its being a legitimate authority, one makes the merit of authority dependent on its achievements – its issuing sound (meritorious) commands overall. Thus explaining the merit of following authoritative commands, in terms of their source in legitimate authority, (itself defined as one that issues sound/meritorious commands) would be circular. Moreover, this explanation itself exhibits a possible incoherence in the Razian view that 1) authority *necessarily claims* to provide subjects with content-independent reasons for action, and 2) this claim, while often not justified, is not conceptually confused. The incoherence here is that authoritative commands are both content-independent and content-dependent: when they are valid, the content-independent authoritative commands are such on content-dependent grounds.[21]

Next, even if one tries to "derive" the merit of performing the requested act from the first potential source of merit mentioned above – that of performing an instance of a "meritorious" act types, one still faces the charge of incoherence. Again, the merit of acting as requested will ultimately depend on the merit of the act-type, and thus be content-dependent after all.

One might try to respond to these incoherence charges thus: they neglect the fact that in the discussed cases Raz's "no direct connection to the action" requirement for CiR has been met, so the above critiques are misdirected. This response should alert us and prompt examining the adequacy of this requirement. The problem is that these seem legitimate critiques: they should be responded to, and not avoided by definitional arguments, based on an unexamined requirement for CiR.

Requirement A anyway seems insufficient to provide a test for content-independence, guaranteeing the distinctness of CiR. To be sufficient, it should be able to specify how "indirect" the connection with merit should be, in order to have content-independence. Is it sufficient that the merit is not grounded in the evaluative properties of the actually required action, but may reside in those of the act-type, or yet a further "distance" from the merit even of the act-type is also necessary? It should also clarify what is meant by "no direct connection" – since "connection" is liable to many interpretations. If this

[21] For such a critique, see Hurd (1999: 80 – 89).

requirement does not provide clear guidance in these two respects (of "how much" indirect, and "what kind" of connection), it may well be inadequate for distinguishing content-independent from ordinary reasons.

Analysing its adequacy in these respects prompts comparing it to the alternative "no dependence on the evaluative properties of the action" requirement. Only after an adequate requirement for content-independence is identified, one could address the above-mentioned arguments against the coherence of CiR concept.

2.1.3. "No Dependence on Evaluative Properties" Requirement

The requirement of "no dependence on the evaluative properties of the action" (B) seems better suited to distinguish content-independent from content-dependent reasons. First, the concepts used are less vague in not appealing to such scalar properties, as connection being more or less "direct" that allow for degrees. Introducing scalar properties in the definition of a concept poses a threat for its correct application. Further, whether sufficiently indirect connection is present, so that one can classify a reason as CiR, may not be independent from one's pre-judgement whether the reason in question is content-independent or not. Such requirement thus cannot serve as the *mark* of content-independence (and cannot provide clear guidance in distinguishing content-independent from other reasons) since its correct application depends on judgements, which already appeal to content-independence. "Dependence" in requirement B is more immediately discriminating: it is an all-or-nothing matter.

Second advantage is that it has clearer meaning than A. As noted, "connection" in A is a relatively "thin" term, not having a clearly specified content, not true to the same extent in the case of "no dependence on evaluative properties" in B. This characteristic of B also restricts resort to judgements, involving an objectionably circular appeal to content-independence, in applying this requirement.

Third, B is connected to the ontological thesis[22] that the validity of these reasons does not depend on the merits of the action they require.

In addition to these advantages, it helps to focus on the *main problems* with content-independence that need be addressed. According to Raz,[23] the main problem for

[22] In section 2.1.1. above were provided some arguments in this respect.

explaining Ci of reasons is that this characteristic of certain reasons (those provided by promises, agreements, mandatory rules, plans for action, etc.) does not seem to cohere with a general point about reasons for action. Normativity, the normative force of reasons (what one ought to do), on this view, is ultimately based on evaluative considerations, on the value of the action (what is good about doing that action). Content-independence also violates transitivity of justification (if A is a reason for B and B reason for C, A is a reason for C) – the reason for having an authority is a reason for the validity of its rules, but is not itself a reason for authority issuing the concrete rules it actually issues rather than others, nor is it itself a reason for doing what those rules demand on a particular occasion. Raz says:

> "The opacity and content independence of rules mean that transitivity [of justification] does not hold. That it is good to uphold the authority of the committee is a reason for the validity of its rules, including the rule that one may not bring more than three guests to social functions of the club. But the desirability of upholding the authority of the committee is not a reason for not bringing more than three guests." Raz (2001: 11)

2.2. The Coherence of CiR: The "Normative Gap" Problem

Thus the main problem for establishing the coherence of the concept of CiR (reasons provided by promises, agreements, adoption of rules, plans for action, and not just by authority) is that their central, characteristic feature – that they are content-independent, contradicts the defining, central feature of reasons for action. This feature is that the normative force (what one ought to do) of reasons, depends on evaluative considerations, on the value of the action (what is good about doing it).[24]

Also, and connectedly, content-independence violates transitivity of justification. Justification is in principle transitive: if A is a reason for B and B reason for C, A is a

---

[23] See Raz (2001a).

[24] Some theorists deny that there is a necessary connection between what one has a good reason to do (good reason depends on the evaluative properties of the action) and what one is normatively required to do. Rationality (what one is normatively required to do), on this view, may bring one to act against good reason Broome (2000). Such position necessarily denies that good *reasons* can ever be content-independent: reasons are based on evaluative considerations only. One can, nevertheless, be normatively required to act in a way, entirely disconnected from evaluative considerations. So, there can be a content-independent *justification* for acting in some ways. Such justification is disconnected from any evaluative reasons for action – one is normatively, rationally required to act in this way. What this position denies is that such a content-independent justification gives one a *good evaluative reason* to act: one is just rationally required to so act.

reason for C. However, justification is not transitive in the case of authority (as well as in the case of the other sources of content-independent reasons for action): the reason for having an authority is a reason for the validity of its rules. But the reason for having authority is not itself a reason for authority issuing the concrete rules it actually issues rather than others, nor is it itself a reason for doing what those rules demand on a particular occasion. Rather, the reason to do what authority directs, is in the "say so" of the authority's rule: it is content-independent.

Thus content-independence introduces what Raz calls a *normative gap* between what one ought to do (the normative force of the reason) and what is good about doing it (the value of the action). How such "peculiar" reasons (characterized by this normative gap) could in principle be valid, needs to be explained in order to maintain that the concept of CiR is not incoherent.

The explanation problem is exacerbated by the fact that the normative gap is not local: it does not appear only at the level of the authoritative reasons for concrete actions. Rather, Raz insists, even the justification of the particular authoritative directives providing agents with content-independent reasons is itself content-independent. This feature is explained by the fact that justification is in principle transitive. This demonstrates, I believe, why Raz's own "no direct connection" requirement is insufficient to distinguish CiRs: it only denies connection between CiR and the concrete action, for which it is a reason. If the justification for the particular directives (being in principle transitive) is itself also to be content-independent, a connection between the reason for the concrete action f and the reason for performing the act-type F is also denied. An explanation for how doing *F* is justified is needed. Appealing to the overall merit of having justified authority does not help. If the justification for doing what authority commands is itself content-independent, this might imply that at the third level, the level of the justification for having an authority, which could issue such content-independently justified content-independent directives, we again have content-independence, and so on. This, of course, is troubling, because the normative gap would not be local, but infinite.

To see why the normative gap would be infinite, consider the converse case. Starting from the other end of the justification chain, if justification is transitive, the evaluative consideration - the ground of the initial reason (the value of having an authority), will be

"transmitted" to the last reason and the normative gap will never open. The ultimate explanation as to why X should do f, will reside in the initial reason. Thus if there is a normative gap, and justification is transitive, then this gap should be *infinite*. There is no point at which it can start to "open", if justification is transitive.

Two conclusions could be drawn from this problem. One is to admit that reasons cannot in principle be content-independent, because such reasons introduce an infinite normative gap. Call this the *incoherence* charge. The other is to conclude that reasons (by definition based on evaluative considerations) are divorced from normative validity (what one ought to do).[25] For such a position, the "normative gap" problem need not be a problem: it is just a fact, about the normative universe we inhabit.

The charge of incoherence should trouble those theorists, who share Raz's position about the dependence of normative validity on reasons, and insist at the same time that CiRs can and do play an important role in any account of authority. The position that justification/validity is essentially or primarily based on evaluative considerations - at the end of the chain of justification, there is some value: the explanation of why one ought to do f is that one has a reason to do f, which is grounded in the evaluative properties of f-ing, - coupled with the point that justification is in principle transitive, is threatened by admitting the validity of CiRs, since they introduce a difficult to deal with within this theoretical framework normative gap.

Discussing the different strategies for dealing with this problem is a task I leave for the concluding part of my thesis. My aim at this point was only to introduce one of the main concepts used in the analysis to follow and to indicate some of its central problems.

Let me now turn to the truly protean concept, charged with doing most of the work on Raz' account of practical authority: the concept of an exclusionary reason for action (henceforth ER).

3. Defining Exclusionary Reasons

As already mentioned, both the conceptual coherence of the concept of ER and its capacity ever to be valid, have been under attack. After defining the concept, I discuss some main arguments against the ER account of authoritative directives. Before going

[25] This is John Broome's position in Broome (2000).

into the details and problems with this concept, let me stress the close interconnectedness of the two elements of protected reasons on Raz's account. Raz explains the plausibility of the CiR concept through that of ER:[26] he believes there are clear advantages to be gained by introducing the concept of ER in understanding such wide-spread normative practices as following mandatory rules, giving promises, obeying authority, making binding commitments, etc, all of which involve CiRs as well. My discussion, then, might necessarily move back and forth between the CiR and ER concepts. To the extent possible, I will, nevertheless, try to separate the issues.

When addressed with an authoritative directive, we feel bound to follow it, disregarding the reasons we have against it. This special character of the authoritative directives is, according to Raz, best accounted for if it is realized that these directives provide a peculiar type of reason. Raz labels it "an exclusionary" (or, as he calls it in some of his texts, "preemptive") reason and defines it thus:

> Def: ERs are *reasons not to act for* some of the *valid reasons* against the commanded action.

The Preemption thesis, stating that when addressed by an authoritative directive, one is to follow it by substituting its reasons for one's own, which are thus being replaced, or excluded, captures this feature nicely.

ERs are also characterised as "excluding by kind, not weight." They may exclude even very weighty reasons of one kind, when they fall within their scope of application and not exclude even trivial reasons of a different kind, if this latter kind is outside their jurisdiction. In addition, as a minimum, commands through ERs exclude considering addressee's own present desires: present desires, no matter how strong, always fall within the scope of application of ERs and are necessarily excluded. [27]

Is the concept of a reason with all these peculiar features coherent?

3.1. The Coherence of the Concept of Second-order Reason for Action

3.1.1. Conformity v. Compliance with Reason

[26] Raz (2001a). I discuss Raz's argument to that effect in further detail in chapter 8.
[27] Raz (1979: 22-24)

ER is a second-order reason - a reason not to act for a reason. The first challenge, accordingly, regards the concept of a second-order reason. The often-voiced dissatisfaction with this concept may stem from the requirement that reasons for action specifically *guide* action: it is not immediately obvious how second-order reasons could do that. Guiding action, it is thought, requires that reasons demand *complying* action. The reason A for X to F seems to require of X to F for that very reason A specifically.

For example, that "it is raining" (A) is a reason for X to take his umbrella, when going out, and X ought (other things being equal) to take the umbrella for that very reason A (because "it is raining"). The second-order reason to F (take the umbrella) for a reason B (that one's mother has said so), on the contrary, also recommends F-ing, without requiring compliance with A (the reason that "it is raining"), but, rather, by requiring compliance with B (the "say-so" of the mother). What may render valid this second-order reason to F for the reason B, is that F-ing for the reason that B is justified. Achieving better conformity with A, without direct compliance to A but by compliance to B instead, may be such a justification.

To conclude, Raz's case for the plausibility of the concept of second-order reasons for action rests on maintaining that it is not generally true that reasons for action are reasons for compliance only. For it might be self-defeating to try to act on a reason by trying to comply with it specifically. For example, respect for the moral law, providing one with a reason to love one's children for their own sake, would be defeated by loving one's children out of compliance with the moral law. The way to respect the moral law is not by complying with that reason directly, but rather by complying with the reason to love one's children for their own sake.

The main point to be stressed here is that, according to Raz, what ultimately matters is conformity to reason, even if it is achieved for some other reasons (by complying with those other reasons, say). This position rests on Raz's view that (1) since there are many ways to satisfy a reason and it is not wrong to act only on some and not all of them, and (2) since there are clear advantages in terms of improved conformity to be gained by not always requiring compliance to the main reason (anyway just one of the many ways of bringing conformity to the main reason), it is better to do so.

3.1.2. The Special Case of Agent-relative Deontological Reasons

This position is plausible, at least when certain agent-neutral reasons are concerned. The problems with it start in the case of deontological agent-relative reasons.[28] If there are valid deontological agent-relative reasons for action (and I will say nothing here on the issue of their validity), one may plausibly claim they require strict compliance rather than conformity. Thus, if one has a deontological reason not to kill innocent persons, this reason requires compliance, rather than simply conforming action. I do not find Raz's argument to the contrary convincing. Since I believe this issue to be important for evaluating Raz's account of the concept of authority, let me discuss it in somewhat more detail.

Raz's argument[29] for "conformity only" in the case of omissions - he does not consider the special case of deontological reasons for action specifically, though the example he uses can be interpreted as involving precisely such reasons - starts by pointing out that the "best mental background" for omitting to commit a wrongful act is that the thought of committing it never crosses one's mind. When one does not kill, one *does not act for any reason* in continuously omitting to kill. Thus the action (omitting to kill) conforms to the reason one has for it, without being the worse, rather, being the *better* for not directly complying with that reason. The increment of "better-ness" here is added by the fact that a person omitting to kill without acting for the moral reason, prohibiting killing, is more admirable than the person who acts for this reason specifically.

This argument is an insufficient support for the "conformity to reason only" argument in the case of deontological reasons. It does not show that conformity to a deontological reason not to kill would have been achieved, for example, were one instead to comply with a non-moral, prudential reason (fear of being caught, punished, etc.), bringing about the same action. It only shows that there need not necessarily be direct compliance in order to have *praiseworthy* conformity to reason.

Further, it does not establish that the conformity to the reason against killing could be achieved by acting on some *other reason* instead. Thus the case of deontological reasons contradicts Raz's statement that

[28] I provide a definition of agent-relative reasons for action, as well as an argument when some such reasons can be valid, in the next part of my thesis.
[29] Raz (1990b: 181).

> "There is no loss, no defect, no blemish or any other shortcoming, in conformity with reason achieved not through compliance with it, but for other reasons" Raz (1990b: 182).

It only shows that acting for *no reason* here is more praiseworthy than directly complying with it. This is hardly sufficient to establish the case for a second-order reason for action - the action in Raz's example is to be done for *no* reason at all. Second-order reason to do or not do an action is a reason to do/not do it *for certain specific reasons* - where these latter reasons are explicitly stated (since this is what gives the content of the second-order reason) - and not for doing/not doing it *for no reason* at all. To clarify the point: notice that though Raz is right that compliance with the reason not to kill is not always required in order to have conformity to that reason, if the omission (not killing), is to be done for *some* reason at all (i.e. if one has a second-order reason against killing), this latter should be specifically that it is wrong to kill.

In short, conformity to one's reason not to kill cannot be achieved by acting for some other reasons (complying with some other, prudential or not, reason), though it still can be achieved through failing to act for any reason at all.

This important conclusion concerns the success of Raz's account of authority in terms of protected reasons for action, comprising a second-order reason for action. If subjects do indeed have valid agent-relative deontological reasons, the claim authority necessarily makes to always determine when those reasons should be preempted and thus not acted directly upon, may not be justifiable. This is especially relevant when discussing the political and legal authority's claim to comprehensive normative supremacy: a task I undertake in the second part of my thesis.

For the purposes of the current chapter, it is important to stress that there is nothing, which a priori prevents the possibility of having reasons to act for a reason. Reasons for action are generally (barring the special case of deontological reasons) reasons for conformity with that action only.

3.2. The Coherence of ER as *Negative* Second-order Reason[30]

[30] The discussion in the following two sections has been influenced by Edmundson's review of Raz's Postscript (1990b) devoted to the issue of ERs, in Edmundson (1993).

One important characteristic of ERs that needs clarification, is that these reasons *exclude* acting on valid reasons, where the latter retain their validity (are not cancelled).

An even more important, central feature is that ERs always win in conflict with the valid first-order reasons falling within their scope of application. This does not imply, according to Raz, that ERs are absolute reasons for action, nor that they have weight that can never be outweighed by countervailing considerations. Rather, their most peculiar feature is that without being absolute reasons or reasons with unsurpassable weight, ERs always win in cases of conflict with the first-order reasons within their scope of application, where their victory does not depend on the relative weights of the competitors. The conflict is resolved not in the normal way conflicts between first-order reasons are resolved - by balancing their respective weights. Rather, ERs win by kind, not weight. It is this latter characteristic of ERs that accounts for the sense in which the protected reasons provided by authoritative directives are *binding*: that they win does not dependent on the relative weight of the countervailing considerations.

An explanation both for the first feature (denying action-guiding role to valid reasons) and for this latter, quite mysteriously sounding feature (ERs, though reasons, win by kind, not weight over other reasons) Raz offers in the Postscript to the second edition of his *Practical Reason and Norms*.[31] His explanation proceeds from what is a plausible resolution strategy for partial conflict of reasons. Before discussing the success of this explanation, let me outline in some detail what is precisely the problem.

A straightforward challenge against the concept of an ER as a second-order negative reason is that it is not a reason *for action*, properly speaking, since its meaning is that one has a reason *not to ground* one's action on certain valid *reasons*. ER is not simply a reason *to believe* that one is mistaken in what one takes to be one's first-order reasons for action. Even though ER is not an epistemic reason (for belief), it is not an ordinary reason for action either. It is *not action-guiding* in a further sense, going beyond the sense we investigated in the case of reasons for action generally (for Raz, recall, reasons for action are only legitimate *direct* guides for action, and do not generally require strict compliance in acting *for* them).

[31] Raz (1990b).

ER cannot be action-guiding (neither in the weak sense of legitimating an action that could in principle be done for other reason as well, nor in the strong sense of strictly requiring an action in direct compliance with it), since it *does not in any way point in the direction of any action*. "Pointing in the direction of an action" is a primary function of reasons for action. Raz at times seems to recognize that ERs are not reasons *for action* proper:[32] they are reasons that license or forbid grounding one's action on certain *other reasons*.

Further, ERs require "negative" compliance specifically: one should *not do* whatever action one has reason to do, *for* certain specified reasons, though one could still do the action for other, not excluded reasons. What is excluded is not the action itself, but doing it for certain reasons only. ERs, then, in distinction to the first-order reasons, demanding conformity only, specifically demand compliance – a negative compliance.

The problem, then, is to explain how a reason which is not action-guiding *stricto sensu*, can nevertheless, require such strict compliance.

The ER concept goes a step further than the concept of CiR. The latter reason is certainly peculiar in presumably guiding action, without specifying the evaluative characteristics of the required action (which is the normal way for guiding action). Nevertheless, it still points by itself in the direction of a *determinate* action, ER, however, cannot even determine *what* does it require, what the action to be performed is, without essentially referring to other reasons. In short, ERs do not themselves guide action, since they only indicate that whatever action is to be done (on considerations other than the exclusionary reason), it should not be done for certain valid reasons. The problem is to explain how considerations, which do not themselves recommend any action, and thus have no direct or indirect action-guiding role can, nevertheless, deny a valid reason its action-guiding role.

3.2.1. The Partial Conflict Resolution Argument

The explanation Raz gives is that ER can deny the action-guiding role of a valid reason as an outcome of partial conflict between valid reasons. Since ER normally only partially

[32]See Raz's Postscript (Raz 1990b).

conflicts with the reasons it excludes, so that the excluded reasons can be conformed to in some other way than by complying with them (thus both the exclusionary reason [requiring strict non-compliance to the excluded reasons] and the excluded reasons [which could still be indirectly conformed to] could be satisfied), then ER could always win. Were it the case that the excluded reasons win, the ER would be completely frustrated, while in the case of a victory of the ER, the excluded reasons need not be frustrated at all. The resolution of the conflict between ER and the excluded reasons relies on the general consideration governing cases of partial conflict resolution: when both sides could be satisfied, it is always better to do so.

The advantage of Raz's solution is that the fact that ERs always win in conflict with the excluded reasons, is explained, without relying in any way on considerations of weight. The way the victory is guaranteed is by following the general logic of partial conflict-resolution (if both sides could be satisfied, because they do not exclude each other, it is always better to do so).

This non-weight balancing conflict resolution strategy in the case of ERs is in stark contrast with the way the conflict between first-order reasons is solved. The latter always relies on establishing the greater weight of the victorious reason. The non-weight balancing conflict resolution strategy has the advantage of better accounting for the way we treat the reasons, provided by mandatory rules, decisions, authoritative directives, promises, etc. We do not think they require balancing against the other reasons we have, but, rather, they always win in conflicts with them, irrespective of their relative weights. This solution, moreover, allows for considerations, which do not have direct action-guiding role, to deny certain valid reasons their action-guiding role. The capacity of ERs to do so is, to repeat, an instance of the logic that governs partial conflict resolution, not relying in any way on the relative weight of the reasons.

A counter-argument to this elegant solution to the problem of the coherence of the ER concept is advanced by William Edmundson.[33] It simply challenges Raz's insistence that in cases of partial conflict between reasons (either between first-order reasons alone, or between first-order and exclusionary reasons) "the question which is the more important

[33] Edmundson, (1993: 329-343)

reason does not arise."[34] Recall that for Raz if both of the conflicting reasons can be satisfied, this should be done (the more reasons satisfied, the better) and the resolution need not depend on the relative weights of the two reasons.

However, Edmundson argues, one has reason to conform with greater number of reasons, rather than with fewer, only when "other things are equal."[35] The other things need not be equal if the greater number of reasons were not as weighty as the fewer: to determine this, however, one need enter into considerations of weight. For example, if one has a reason to eat, one has a reason to conform to this reason, which can be done by eating slowly, quickly, for pleasure, etc, as well as in any conceivable way that would satisfy the reason to eat. In case of a conflict with an exclusionary reason not to eat quickly, the obvious solution is to require eating in any other way except eating quickly (since in this way neither the first order reason to eat, nor the exclusionary reason not to eat quickly will be frustrated). If the reason to eat is specifically to eat quickly (since one is famished, and eating quickly satisfies the reason to eat when famished to a greater degree), the resolution of the conflict with the exclusionary reason not to eat quickly will depend on the relative weights of the two partially conflicting reasons. The conflict is still partial, since one can, in principle, satisfy the reason to eat, when famished, in some other way then by eating quickly. However, it is not immediately obvious, that since one can satisfy this reason in other ways, one always should. It might be more important to satisfy the reason to eat to a greater degree, since it is urgent, while frustrating the purported exclusionary reason is relatively less important.

More interestingly, in cases of conflict with ERs, which can be expected to reduce the chance of conforming to the ultimate reasons, this chance creates a second-order positive reason to act on the balance of all first-order reasons. The presence of this second-order positive reason should not be allowed by Raz to turn the partial conflict into a head-on one, to be decided by the weight of the two second-order reasons. Were Raz to allow this, there would be no merely partial conflicts: since one always has a second-order reason to get the balance of the first-order reasons right. If having partial conflicts (presumably implied by the absence of second-order positive reasons to act on the balance of all first-

[34] Raz (1989:1167).
[35] Edmundson (1993: 339). In the discussion below, I closely follow Edmundson's arguments.

order reasons) depended on the likelihood that the excluded reasons will, as a matter of fact, be satisfied, ERs would loose their distinctiveness. Having ERs would be indistinguishable from having first-order reasons, which happen to outweigh all the conflicting reasons. To save Raz's account of ERs, it needs to be maintained that the resolution of partial conflicts between ERs and second-order positive reasons to act on the balance of all reasons, is to be done irrespective of weight. The plausibility of precisely this claim is challenged by Edmundson and has not been securely established.

Thus, if Edmundson is right that even in partial conflicts the question of the relative weights of the reasons may need to be and often is considered, Raz's way of dissolving the paradox of "having reasons not to act for some valid reasons", is unsatisfactory.

To securely establish his case for ERs as always winning over first-order reasons within their scope of application by kind, irrespective of weight, Raz has to go back from this formal in character argument from partial conflict resolution strategy and resort to his other arguments.

One further, structural in character argument relies on establishing that in order to make the practical difference to one's reasoning they are meant to make, authoritative directives (as well as promises, mandatory norms and rules, decisions, etc.) should exclude acting on (some of the) considerations, hostile to the presently required by the directive course of action. Raz's Preemption thesis, which captures precisely this characteristic, is the main structural thesis of the Service conception. It flows directly and draws its plausibility from the two moral, normative theses of this conception: the normal justification and the dependence theses. The fact that the question about the coherence of the ER concept cannot be resolved without ultimately resorting to normative arguments confirms Raz's position about the close interconnectedness of conceptual and normative issues.

3.3. Problems with Weight and Scope

Let me point to a different problem with ERs and their coherence, which has not drawn the attention it deserves. It again pertains to their status of reasons, but it concerns their weight - do they have it and how is it determined. The discussion on ERs has predominantly been focused on the issue of how can they always win in their conflict

with first-order reasons within their scope of application, if their victory does not rely on their weight. Whether those reasons do have weight and how it is determined, has not been an issue. I find this puzzling, given that the issue of weight has bearing on one of the central features of ERs – they have limited scope of application. Precisely this feature, recall, allowed Raz to use the partial conflict resolution argument in support of their coherence: were they to be absolute, the conflict could not be partial.

The importance of the issues of the weight and the scope of ERs, however, goes much beyond the support for this particular argument. Crucially for the purpose of Raz's analysis of authority and the success of his response to the critiques of philosophical anarchists, ERs need to have a *limited* scope of application. If a reason with no determinate weight, nevertheless "excludes" and thus wins over the countervailing reasons, the explanation might be that it is *absolute* in its scope of application, as well as in its weight. It is otherwise incomprehensible how it could always win irrespective of weight.

So, it could be argued that the explanation of why ERs always exclude the reason falling within their scope of application, is that there are especially weighty reasons for this: it is the weight of these reasons which determines the scope of application of ER, within which they always win. What is needed is an explanation how the weight of the reasons for singling out ERs, determines the weight and the scope of ERs themselves. Notice that ER with an *indeterminate weight and scope* of application is anyway incomprehensible: it is then indeterminate whether it does or does not exclude any reason, since it is indeterminate whether the latter falls within its scope.

It is important to stress at this point that a solution of this problem with determinacy (already alluded to above): "they are absolute in weight and scope" and therefore one need not bother with exactly determining their weight and scope, is a non-starter. Were they to be absolute, there could in principle be no legitimate authorities: the claim that one has to obey an alleged authority "come what may", could not be rationally justified. This is the philosophical anarchist's challenge. If ERs could be absolute in their application, the problem of precisely determining their weight and scope would not be on the agenda: irrespective of their weight and scope, they would always win over

conflicting reasons. Raz, for the just indicated reasons, claims ERs are not absolute, and do have determinate weight.

There is no obvious answer in Raz's texts, however, not only to the question of how the determinate weight of ERs is established, but also to the more basic question of whether and in what way ERs have weight at all. If it is ambiguous whether and how ERs can have weight, it is also ambiguous whether they are reasons for action, since reasons for action necessarily have weight.

In response to critiques[36] Raz maintains[37] that ERs have weight, and thus *can in principle be compared* with the first-order reasons, but often are *not* so compared, since comparing reasons is not what we do, when we are faced with authoritative directives, mandatory rules, promises, etc. (presumably involving such ERs). We, instead, reason along the lines of the partial conflict resolution strategy: the more reasons we can satisfy, the better. Chaim Gans agrees that weighing the countervailing considerations is not what is involved in having reasons, provided by promises, mandatory rules etc., which do conflict with our first-order reasons. He nevertheless contests Raz's claim that ERs are involved. Rather, we might instead have a case of incommensurable reasons. On of Gans's arguments[38] challenges the evidentiary (phenomenological) test Raz advances for establishing the plausibility of the ERs account of mandatory rules. This test draws on the fact that when faced with a conflict between the balance of our first-order reasons and the reasons provided by the authority, etc., we feel a special *unease* in resolving it in favour of the latter.[39] This unease, according to Raz, shows that we do not consider that the former reasons are defeated: if they were defeated, there would be nothing to be uneasy

[36] Gans (1986)

[37]This presentation of Raz's argument for the superiority of his account of authoritative reasons in terms of ER over Gans's incommensurability account of the reasons provided by authority, mandatory rules, etc., can be found in Edmundson (1993).

[38] Similar points against the phenomenological argument for ERs are raised by Richard Flathman (1980), and are extensively discussed by Moore (1989).

[39] "When the application of ER leads to the result that one should not act on the balance of reasons, that one should act for the weaker rather than the stronger which is excluded, we are faced with two incompatible assessments of what ought to be done. This leads normally to a *peculiar feeling of unease*….These two types of situation provide *the test case* for the presence of exclusionary reasons precisely because it is in these situations that their presence makes a difference to the practical conclusion." Raz (1990b: 41, emphases added)

about. Rather, this unease shows that they are *excluded* from affecting the outcome of the deliberation as to what one ought to do.

Gans, however, shows that the phenomenological test does not unequivocally demonstrate the plausibility of the ER account, since an equally plausible interpretation of this feeling of unease is that a conflict of incommensurable orders of (just first-order) reasons for action is involved. Raz's response is that though the two orders of reasons (first-order and exclusionary) are in principle *comparable* (which contests Gans's claim that a case of incommensurability and not ERs may as well be involved), they often are *not so compared*. Such is the case of acting to fulfil one's promises, to obey authority, i.e. in cases of ERs.

My contention (and the reason I brought here the debate between Gans and Raz) is that the comparability claim goes against Raz's concession,[40] that ERs are not reasons *for action*, properly speaking: they are reasons against acting *for certain reasons* only. If ERs are not typical reasons, they might indeed not be comparable to the first-order reasons for action, since the basis of the comparison is the weight of the reasons. Does the concession that ERs are not typical reasons for action imply that those reasons are not comparable to the first-order ones? This would be so, if the explanation why ERs are not typical reasons has to do with them not being capable of having determinate weight. If ERs do not have determinate weight, however, it is unclear how the conflict between them and the positive second-order reasons is resolved: Raz agrees that in cases of such head-on (rather than merely partial) conflict, the resolution is on the basis of their relative weights.

Raz's discussion of the problems with the coherence of the concept of ER concentrates on a different issue than the one that interests me here. The focus is not whether ERs themselves are reasons for action (which presumably have determinate weight) but, rather, whether there can be valid ERs, which require not acting for certain valid reasons.

> "Exclusionary reasons are reasons not to act for certain reasons. This gives them the appearance of paradox…After all, reasons are there to guide action. Surely, there cannot be reasons for not being guided by reasons, whose very nature is that they should guide action. The argument [above] helps dispel the air of paradox. It shows that reasons are merely legitimate guides. One does not have to

[40] Raz (1990b), Postscript.

> be guided by them. Other things being equal, so long as one conforms with them, there is nothing wrong with one." Raz (1990b: 183)

Raz here does not address the issue whether ERs are themselves reasons for action, which presumably *themselves need to guide*. His claim that reasons need not necessarily guide action is not good in the case of second-order reasons: ERs as second-order reasons require (negative) *compliance* specifically, not simply conformity. The argument here establishes that first-order reasons are only legitimate guides to action and need not be complied with, in order to guide it: conformity is enough for that. Thus if by complying with other reasons (ERs) they could still be conformed to, there is nothing paradoxical in those ERs.

What this reply does *not* establish, however, is that ERs themselves as reasons *for* action are not paradoxical. It does not establish that ERs themselves can in principle guide action - either through conformity or through compliance with them. This latter also should worry the theorists doubting the *coherence* of the concept of ER. This worry is connected to the one, raised by the concept of CiR, but goes deeper. The question is not simply how a reason, which does not point to the good of performing an action, can guide action. Such a reason at least can by itself have *determinate content*, and the required action - specified by reference to CiR alone. Rather, the worry here is that a reason, which relies for its specification on altogether independent reasons, and thus the required action has no content, identifiable independently from reference to such reasons, cannot possibly be said to guide action either through conformity or through compliance. ERs are not reasons *for* action: they do not point to the desirability of performing certain actions, but rather, point to the desirability of *not* acting on some independently from the ERs specified first-order reasons.

My claims are supported by Raz's text. He defines ERs as "reasons for acting, the full specification of which essentially refers to other reasons."[41] Importantly, this definition does not sit well with an argument Raz himself advances against the critique that his ERs cannot be distinguished from "canceling reasons."[42] Raz responds to this critique by denying the "canceling reasons" the status of *reasons*. His argument for this is precisely

[41] Raz (1990b: 185)
[42] Raised by Moore (1989)

that by their nature they are *not action-guiding*. He prefers to refer to them as "canceling conditions" instead. Raz's argument for the distinction between ER and canceling condition is precisely that "canceling conditions" necessarily "*relate* to the reasons they cancel,"[43] while ERs are presumably not such. This is so, because ERs "are *essentially independent* considerations which point to the desirability of the non-performance of the action."[44] As we saw, Raz also defines ERs as "reasons, the full specification of which *essentially refers to other reasons"* (emphasis added). The definition of ER here is close to the characteristic feature of a "canceling" condition mentioned above. The basis for distinguishing canceling condition from ER - the possibility of specifying the latter but not the former without essentially referring to other reasons for action - disappears.

The upshot of all this is that it is not clear in what sense ERs have weight. I take it that, generally, the weight of a reason is a function of the desirability of the action it recommends. Since ERs do not by themselves recommend any action, properly speaking, nor are they action-guiding neither directly nor indirectly, it could be argued that they do not have weight, or at least it is not clear how this weight is determined, if they have one. This simple argument, however, goes against Raz's insistence that ERs have weight, even if indirectly[45] attributed.

The considerations I brought forth are not sufficient to establish that ERs cannot have weight. They do, nevertheless, raise serious doubts in this respect. Next, the challenge against Raz's claim that ERs have weight, so can in principle be compared to other reasons, but often are not so compared, does not directly threaten the *coherence* of his *conceptual* analysis of orders in terms of ER. There is nothing unintelligible in the way they can always win in cases of partial conflict with first-order reasons. However, it threatens the plausibility of this analysis. Thus, I find wanting his reply to Gans's

[43] Raz (1990b: 188, emphasis added)
[44] Raz (1990b: 188, emphasis added)
[45] For this claim, consider Raz's discussion of rules and mandatory norms as providing ERs: "Rules are not ultimate reasons. They have always to be justified by more basic considerations. This is a result of the fact that norms are exclusionary reasons. A reason not to act for a reason cannot be ultimate. It must be justified by more basic considerations. Furthermore, rules normally represent the result of considering the application of a variety of conflicting considerations to a generic situation. This explains why they are not ultimate. It also explains why the reasons for the norm are not always obvious from the formulation of the norm… Since a norm is the outcome of the requirements of various conflicting values *it does not carry its desirability on its face*. It simply states what is required, of whom and when, but it does not always do so in a way which makes obvious the reasons for the requirements." Raz (1975: 76).

challenge: he might be right in insisting that what Raz describes as cases of ER, are rather better understood as cases of incommensurability.

This, however, is not the only, nor is it the main conclusion to be drawn from the discussion above. More importantly, one of the main elements of Raz's position on the possibility of having *valid* ERs can also be challenged. The idea of ERs having certain *limited scope*, only within which are the countervailing reasons validly excluded, might turn out to be unintelligible. The problems with the limited scope of ERs may come from both sides of the debate.

In the case Raz is right that the best interpretation of what reasons we have when faced with authority is in terms of ERs, given the doubts already raised concerning the weight of those reasons, one may wonder whether and how such reasons with "indeterminate" weight could have *limited scope* of application. The exclusion by such reasons, contrary to what Raz claims, by not depending on the weight of these reasons, might as well be blanket, indiscriminate, in a word – absolute.[46] Arguably, there is a connection between a reason not having a determinate weight but nevertheless excluding the countervailing reasons, and its having unlimited, absolute scope of application. The idea of ER with an indeterminate scope of application, it was already noted, anyway seems outlandish. Were it to have indeterminate scope, it would be indeterminate whether it does or it does not exclude any reason from affecting the outcome of the deliberation. Exclusion requires determinacy as to what is excluded: it could either be a clearly specified (by the limited scope of the ER) exclusion, or it could be a blanket exclusion (if ER has unlimited, absolute scope).

If, on the other hand, the cases of what Raz describes as involving ER are better understood a la Gans as involving incommensurable reasons (thus accounting for the feeling of unease when acting on the reasons provided by the authority), then the reasons, provided by authority, etc., would again need to be *absolute*. Otherwise those reasons could not *always* win in cases of conflict with the first-order reasons against the commanded action. To see why, notice that since the orders of reasons are, by assumption, here incommensurate, i.e. cannot be compared, if one of the orders is to win,

[46] This would, incidentally, threaten Raz's response to the anarchist challenge. It could only be rational, if at all, to comply with an authority, when the directives it issues are with a limited scope of application.

it should be absolute, limitless – to be always complied with no matter what. Here it is the incommensurability of the two orders of reasons, which does not allow for having "limited" scope of application of the authoritative reasons. This is so, because if in order to decide whether the authoritative reason wins over the incommensurable conflicting reason, one is to see whether the latter falls within the scope of the former, this means that the two reasons are in fact comparable.

The above-indicated problems with the limits on the scope of application of authoritative reasons, recall, pose a threat to their validity as well. Obeying authority issuing absolute exclusionary reasons, to be complied with under all conceivable conditions, cannot be rationally and morally justified. Notice the grand implications if such absolute reasons cannot in principle be valid – then in principle there could not be valid promises, legitimate authorities, valid agreements or mandatory rules. This is so, since one cannot rationally justify[47] complying with an agreement, keeping a promise, following a mandatory rule "come what may" - under every conceivable circumstances (which is required, if the reasons provided by promises, etc., were to be absolute).

This is a conclusion Raz would like to avoid: so, further arguments are necessary to establish that ERs are reasons with determinate weight and limited scope of application, which nevertheless always win in cases of conflicts with the reasons falling within their scope of application.

4. Conclusion

In this chapter I have introduced and discussed the main elements in Raz's account of the concept of practical authority. I have laid out the structural features of this concept, and discussed some of the theoretical problems with its central concept of a protected reason for action and its two components – CiR and ER. I have suggested that the distinctness of CiRs is best drawn out if it is characterised by a "no dependence on evaluative properties of the action" requirement. I have also shown how this characterisation better illustrates the "normative gap" problem with the coherence of the concept CiR. I have next identified several problems with the conceptual coherence of the ER component of

[47] This rational justification, when present, would render valid ERs, provided by the valid promises, mandatory rules, authoritative commands, etc.

authoritative protected reasons. I found well-grounded the challenges pressed by Edmundson (against Raz's partial conflict resolution argument for the coherence of ERs) and Gans (contesting Raz's ER explanation of authoritative directives). At the end of this chapter, I advanced my own (not entirely conclusive) arguments why ERs as reasons for action may be problematic. They revolved around the issue of the weight and the scope of these reasons. My main concern was whether the problems with determining the weight of those reasons reverberate on the issue of their limited scope. That these reasons could have a limited scope rather than be absolute and that it be clearly determined rather than indeterminate is important, if legitimate authorities are to be possible.

After discussing these conceptual puzzles, I need to turn to the further, no less important question: what are the conditions, under which protected reasons with their two elements can be valid, and acting on them justified. This is one of the main topics of my thesis: though I necessarily need to start from an analysis of the concepts of authority, authoritative directives, etc., what drives this analysis is the main question about the legitimacy of such authority and the justification for acting on its directives. Thus in the following chapter I introduce Raz's essentially instrumentalist Service conception of authority's legitimacy, identify the main interpretations of its legitimacy test, as well as raise several concerns with its adequacy as a test of legitimacy for a practical type of authority.

Chapter Two

The Service conception of Legitimate Authority: Normal Justification Thesis and the "Moral Duty to Obey" Problem

The task of this chapter is to introduce Joseph Raz's Service conception of legitimate authority, as well as to outline what I believe to be the main theoretical problem with it. This conception specifies under what conditions the claim the law (and the state acting through the law) necessarily makes to possess legitimate authority over its subjects, is justified, and its authority - legitimate.[48] In the beginning of the chapter I briefly introduce the conception with its main building blocks. Next, I provide a detailed discussion of its legitimacy test – the Normal Justification Thesis (henceforth NJT). I identify two serious sources of discontent with it. The first is its ambiguous support for the practical difference thesis: recall that legitimate practical authority makes a difference to how its subjects ought to act, by giving them new reasons for action. Next, I spell out in detail the main problem with NJT: it has serious difficulty accounting for the common sense notion that legitimate authority acting within its jurisdiction, provides its subjects with a moral duty to obey it. Thus, on instrumentalist grounds it may be difficult, if not impossible to explain how authority gives new reasons for action. The real challenge for this account, however, is to explain how on instrumental grounds authority can be a source of moral duty to obey. Does this account have the resources of filling in the gap that opens between the *reasons to obey* and a *duty to obey?*

1. The Service Conception of Legitimate Authority.

1.1. The Three Core Theses

[48] I will, for the purposes of this thesis, disregard Simmons's warning not to conflate the issues of justifying (state is morally permitted to perform its characteristic functions) and legitimating (state imposes duties of obedience on its subjects) the state, Simmons (2001). This way of conceptualising this otherwise very important distinction is not uncontroversial. Allen Buchanan (2002: 694), to name just one example, uses the term legitimacy to denote "morally justified wielding of political power." However, the general difficulty in moral theory, that Simmons's distinction is meant to articulate, that of a logical gap between moral/prudential *reasons* favouring certain acts and institutions, and moral/rational *requirements* or even *duties* of individuals to perform such acts and submit to the commands of those institutions, will be one of the main concerns in this chapter. On this logical gap, and a possible solution to it, see Edmundson (2003: 211-214).

The Service conception states that authority is there to serve its subjects by mediating between them and the reasons that apply to them, when subjects need this mediation. It is a generally instrumentalist type of conception, consisting of two core moral elements, and a structural, formal thesis, directly flowing from them. The moral elements are NJT and the Dependence thesis. The structural element is the Preemption thesis. There is a corollary to this conception as well - the autonomy condition. I discuss NJT - the most important thesis, defining the character of the Service conception by providing its legitimacy test, in the next section of this chapter in detail: its different interpretations with their relative advantages and attending problems. The main ambition of this conception, and its main advantage, if successful, is to dissolve the rationality paradox that plagues the inherited from the tradition, common-sense conception of practical authority: one cannot rationally obey authority. If one obeys an authoritative directive supported by the balance of reasons, one follows reason, and does not obey the directive. If one does obey a directive not supported by the balance of reasons, on the other hand, one obeys the authority, but is not rational – acts against the balance of reasons. Thus the focus of the chapter is on the indirect instrumentalist strategy: NJT, advanced by Raz for the purpose of dissolving this paradox. Raz's success in this respect would be a success in solving the autonomy paradox as well: the latter is just for him an extension of the former.

The question that an account of the legitimacy of authority needs to answer is when are the *protected* reasons for action provided by a practical authority valid, i.e. when is one under obligation to obey authority's commands (that is, if obligation can be conceptualized in terms of protected reasons for action provided by practical authority). The general answer Raz gives is that authority's role is to serve its subjects by mediating between them and their reasons (the reasons that apply to them prior to and independently of authority). More specifically, authority's claim to legitimacy is normally justified when its directives are likely to allow the subjects to better conform to their pre-existing reasons by complying with the directives rather than by acting on (complying with) those reasons directly. This general answer is Raz's famous "Normal Justification Thesis"

> "…the normal way to establish that a person has authority over another person involves showing that the alleged subject is likely better to 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." (Raz 1986:53)

It is the 'normal' because it is not the only one, though it is according to Raz the most common and the most important justification. [49] This moral in character thesis proceeds from a different moral thesis, concerning the types of reasons that should guide authorities in issuing their directives. This is Raz's Dependence thesis:[50] authority's directives are meant to be based on the subjects' "dependent" (Raz's term for subjects' own, prior to authority) reasons, and are meant to reflect their correct balance. The concluding part of the Service conception is Raz's famous Preemption thesis. This structural in character thesis logically follows from both of the two moral theses above: the reasons provided by the authoritative directives are not to be added to the pre-existent reasons, but should exclude and replace (some of) them.

This conception of legitimate practical authority is the most influential reason-based type of justification of political authority to date. On it, the justification for the exercise of political authority ultimately relies on the sound reasons of its subjects, to which it is capable of bringing improved conformity. It is a reason-based, and not will-based type of justification, since whether authority is or is not justified does not depend on whether its subjects agree with and are willing to abide to its orders, or agree that the test of legitimacy thus specified, is met. It is a matter of objective reasons, and it is a matter of objectively improved conformity to those objective reasons, that ultimately justify political authority. The position here is stated dogmatically – for the purposes of mapping as clearly as possible the conceptual ground. In fact, as it will become clear in the process of my exposition, the issue is not that simple. We will see that there are several interpretations of NJT and there is a qualification on this thesis, imposed by the requirements of "the autonomy condition."[51] Nevertheless, it is undoubtedly the case, that

[49] An ambiguity in this definition should be cleared away: what authority is likely to bring about is not improved compliance to subjects' dependent reasons, but their better conformity to them. Recall the discussion in chapter 1 of the importance for the defence of the coherence of the concept of ERs, of the distinction between compliance and conformity to reason. NJT is one of the central elements of Raz's account of authority, where this distinction plays out.

[50] For a discussion, see Raz (1986: 42-53).

[51] This term was introduced into the discussion of NJT by Green (1989: 795, 810).

Raz's account of political authority's justification is a reason-based one in a strong, rather uncompromising sense.

The qualification above states that even when the requirements of the three theses, comprising the Service conception of legitimacy, are met, this is not sufficient to show that the exercise of political authority is justified and the authority itself - legitimate. Thus a further condition to be met is the *condition of autonomy*. It is uncontroversial, that sometimes it is more important for the subjects to decide and act on their own, not guided by authority, than to decide and act correctly (which presumably authority can help them to do by exacting obedience to its directives). The three theses above, then, justify acting on authoritative commands (impose obligation of obedience) only in cases, where it is more important for its subjects to act correctly rather than act on their own. As I will show later in my dissertation, this restriction poses some considerable difficulties in explaining how law's and state's necessary claim to authority can ever be justified. It is of law's and state's nature that they always claim more authority than can possibly be justified, not only because the requirements of NJT are very stringent (after all, there are alternative justifications for authority, and the requirements of these might be met if those of the NJT are not), but also and especially because of the restrictions of the autonomy condition.

The general character of the Service conception of authority, and of its NJT in particular, it was mentioned already, is not only reason-based, but instrumentalist as well. Authority is on the whole justified whenever it is a good instrument in the service of its subjects - helping them conform better to their own reasons. This characteristic also explains how it could be rational to obey authority: whenever the likelihood of successful conformity to those reasons is thereby increased. In this way, the instrumentalist, reason-based Service conception has a clear "rationality" advantage: the paradox of rationality a conception of practical authority seems necessarily to generate is thus arguably dissolved. Further, since one ought to do, according to Raz, what one has most reason to do, the instrumentalist strategy promises to respond to the autonomy paradox as well. Raz's response consists in denying that one has a moral duty to act autonomously – it is one's duty to act on right reason instead. If authority helps one to discharge this duty of acting on right reason

better than if one always acts autonomously, then the responsible course of action is to decide to follow authority: indeed, it is one's duty to follow such beneficial authority.

The theorists that doubt the success of this instrumentalist strategy for dissolving the rationality paradox doubt its success in the case of the autonomy paradox as well – since Raz's solution to the latter is a replica of his solution to the former. Those critics have challenged either certain elements of this conception - one, or all of its moral or structural theses, or they have attacked Raz's strategy as a whole. Instead of rehearsing in a textbook fashion the critiques, arguments, counter-arguments, etc., (hardly a rewarding or enjoyable task), I start my discussion by offering a systematic exposition of the main interpretations of NJT, in the course of which I indicate the main problems, advantages and disadvantages of this central legitimacy test.

2. The Normal Justification Thesis: Interpretations

2.1. Exclusively Substantive or Inclusive? The "Filtering" Role

There is a wide agreement among the theorists that NJT as defined by Raz is an instrumentalist, piece-meal (person-by-person and issue-by-issue) and indirect approach to justifying authority. Authority is legitimate not because of some inherent properties it uniquely possesses, but only to the extent obeying it helps its subjects conform better to their own, independent from the authority reasons: authority is the servant, not the master. The agreement[52] on the interpretation ends here, however.

One central contested issue is whether NJT is an exclusively substantive test of legitimacy, or it could be construed more inclusively, to comprise some procedural concerns as well, arguably also pertaining to the legitimacy of an authority. Discussing this issue is important for evaluating the prospects of NJT and the Service conception of legitimacy more generally as an adequate conception of the legitimacy of democratic authorities. The procedural dimension is of central concern in a democratic type of authority, and a conception of legitimacy that downplays the importance of the procedural dimension could hardly qualify as an adequate conception for this type of

[52] Some contest even these central points. Christiano (2004: fn. 14 at 278), for example, does not believe the instrumentalist interpretation of NJT is necessary. He does not, however, elaborate on this.

authority.[53] Let me note that the procedural dimension is important not only if a proceduralist account of democratic authority is accepted. Even a substantive conception of democracy has to account for the legitimacy of the main mechanism for decision-making in a democracy – the majority procedure. After all, we consider legitimate the concern within a democracy with how, through what procedure have the collectively binding decisions been reached.

Thus, there would be a serious problem with the Service conception (and NJT in particular as its central moral thesis, itself responsible for the character of the conception), if it could not accommodate these procedural concerns. Thus an often voiced recently criticism of the NJT as a plausible test for the legitimacy of political authorities[54] is that because of its substantive character, it cannot account for the importance of the procedural aspects of legitimacy, especially prominent in accounting for the legitimacy of modern liberal-democratic political authorities. It has been suggested that a way to circumvent this problem is to extend the Service conception to cover procedural considerations as well.

Thus, let me distinguish between two interpretations of this conception: an inclusive and an exclusive one. On the inclusive conception, NJT – the main moral thesis of the conception, can accommodate without significant changes procedural concerns as well.

One possible way to inclusively interpret NJT is to try to exploit the distinction Raz, following Derek Parfit, introduces between action-reasons and outcome-reasons.[55] The value of acting on action-reasons is in the intrinsic value, residing in the *performance* of the action, not in its outcome, as on the outcome-reasons. If improved conformity to action-reasons may be thought to meet the requirements of the NJT, then NJT may be interpreted as permitting procedural concerns to affect the outcome of this legitimacy test. This seems a promising route, since it might show that some procedural concerns could be met by NJT, without necessarily violating its general instrumentalist and broadly substantive spirit – of

[53] Concerns in this regard are raised famously by Waldron (1999) and Waldron (2001). For an argument against the adequacy of Raz's conception of authority, and the NJT in particular, for a democratic type of authority, see Hershovitz (2003).

[54] Apart from the already mentioned arguments in Waldron (1999) and Waldron (2001), the issue has been discussed by Shapiro (2002a), Hershovitz (2003), among others.

[55] On this distinction, see Raz (1986: 279) and Parfit (1984).

bringing improved conformity to whatever reasons apply to their subjects independently of authority.

This route is tempting for a Razian, since it does not require any substantial revision of NJT. I suspect, however, that it might nevertheless be blocked. Bringing improved conformity to action-reasons may be excluded from the scope of legitimate authoritative action. The reason is straightforward: I cannot satisfy (let alone bring improve conformity to) the action-reason I have to myself do my maths assignment, for example, by letting someone else tell me how to do it (by letting him solve it for me). I might, though, improve my conformity to my outcome-reason to reach a correct solution to the same maths problem by submitting to the authority of a mathematician.

This asymmetry of action-reasons and outcome-reasons with respect to authoritative directives is due to authoritative reasons' exclusionary aspect (recall the discussion from the first chapter). This exclusionary aspect is reflected in the Service conception's Preemption thesis. When authority issues its directives, the subject is to take them and not his own first-order reasons as guides for his action: he should act on the authoritative reasons and not on his own, since these latter have been replaced. This is especially significant, if it is the case that performing the action for which one has an action-reason, has an intrinsic value only if the agent performs this action *for this reason* alone – the performing itself has *intrinsic* value, and not for other reasons - performing is *instrumentally* good for something else. Since the presence of authoritative directives is meant to preclude agent's acting *for his own reasons*, the intrinsic value of acting on action-reasons may be[56] lost. This is due to the effect of the Preemption thesis: it rules out *direct* action on the first-order reasons one has. Thus, if one has an action-reason, conformity to it could not be improved by not acting on it. If this is so, action-reasons cannot be covered by NJT. In this way, the prospect of accommodating some kind of procedural concerns – pertaining to the *way* certain result is achieved, (say, by permitting acting on action-reasons directly) rather than the result itself, within NJT, is closed.

[56] I say, "may be" and not "is," since the conclusion here depends on the truth of the conditional, stated in the preceding sentence. It is conceivable that performing the action required by the action-reason "to sing," for example does not lose any of its value if the act of singing is done not for its own sake, but for authority's sake. Nevertheless, though conceivable, this is not plausible. Notice that if action-reason is interpreted to permit performing the action for any reason, the distinctness of this reason – that there is an *intrinsic* value in acting on it, is lost, swallowed by the consequentialist vacuum cleaner with its instrumentalist and maximizing logic. I will come back to these issues at several points in this thesis.

The second prospect for an inclusive interpretation of NJT is no brighter.[57] One might try to argue for including a proceduralist element in NJT by noting that if one has some independent reasons to comply with the results of certain decision-*procedure* (say, democratic majority rule), complying with them will bring improved conformity to these independent reasons. On the face of it, this satisfies the test of legitimacy of NJT. However, it renders NJT empty. Thus the first objection to this interpretation is that NJT would be an empty box, were it to accommodate the results of any test of legitimacy (democratic proceduralism included) that turns out to be true: since we always have (a second-order) reason to do what the right test of legitimacy says we should. The second objection is that Raz does not intend NJT as such an inclusive test of legitimacy.[58] Hershovitz thus rightly (I believe) argues that were Raz to intend his legitimacy test to be interpreted in this inclusive way, he would have stated the autonomy condition as an extension of NJT, rather than as its exception. For Raz, while we do indeed have reason to value our autonomy, it is one thing to act so that to bring improved conformity to one's reasons, and it is another to bracket this concern and act autonomously even at the expense of failing to improve, or even worsening one's conformity to reasons overall.

The above considerations argue for describing Raz's own interpretation of NJT as *exclusively substantive*: good outcomes, irrespective of the process that yielded them, is what confers legitimacy on an exercise of authority. Notice that Raz does not claim that NJT, when met, is sufficient for establishing the legitimacy of a political authority. Even when the autonomy condition also is met, this does not necessarily establish authority's full legitimacy, since there might be other considerations, not related to correct outcomes or respect for autonomy, that pose a challenge to it. Raz believes, however, that NJT (together with the other two theses), as constrained by the autonomy condition, are always necessary and normally a sufficient test for legitimacy.

Notice also that the exclusively substantive interpretation of NJT does not exclude support for democratic procedures if they are more likely to turn out to be overall substantively beneficial as a result. However, the support for any particular type of procedure is entirely *contingent* on its capacity to yield *correct results*, and not on its

[57] This suggestion is discussed and refuted by Hershovitz (2003).
[58] For a clear statement, see Raz (2003: 216).

independent from the result, inherent properties of being a fair, respectful, etc., procedure.

A student of authority with Razian affinities, who wants to offer an account of the legitimacy of democratic authority of a strong type, should not despair at this point. Even when she is not satisfied with the contingent place attributed to democratic procedures on NJT, she could still hold NJT (together with the other two theses of the Service conception) to be a necessary condition for the validity of any justification of authority, democratic included. Their role will be in filtering out certain accounts of legitimacy: thus this is a third, "filtering" interpretation of NJT[59] (or the Service conception more broadly).

Contrary to what might be expected, I think the capacity of the Service conception to play a "filtering" role for any plausible conception of legitimacy, does not mean that the Service conception is not a distinct conception of legitimacy, competing with other accounts. Only being a distinct account, with an exclusively interpreted test of legitimacy, could it serve the filtering role the Razian democratic theorist would like it to play. Were it to be compatible with all accounts, it could not play such a filtering role. Thus, I suggest to distinguish the inclusive interpretation of NJT dismissed above from this latter "filtering" one, taking NJT as a necessary condition for any adequate account of the legitimacy of authority, democratic included.

2.2. Objective Only or a Subjective Element as Well?

A different issue is whether the justification for holding one to have authority over another, is in terms of the *objective reasons for action* of the latter, to which an objectively improved conformity is as a *matter of fact* brought about, or rather, in terms

[59] Besson (2005) argues along these lines for NJT as a "filter," screening potential justifications and establishing a criterion for their legitimacy. Hershovitz (2003) takes what Besson believes to be the opposite position, interpreting NJT and the service conception exclusively - as a particular conception of legitimacy, distinctly substantive in character, and thus in competition with procedural ones. It seems to me incorrect to juxtapose the "exclusive" and the "filtering" interpretations of Raz's conception. Indeed, I think Besson's critique of Hershovitz's position on NJT exclusively interpreted, is misdirected: he seems right that NJT should be seen as a distinct test of legitimacy, competing with other such. Only being such a distinct test, could it serve as a filter for screening *out*, excluding potential justifications. It is in this way only that it could perform the function Besson rightly attributes to it. All this said, I gratefully acknowledge my debt to the discussion in Besson (2005): it will play out in the concluding part of my thesis.

of one's *subjective reasons for belief* that such conformity is *more likely* to be brought about.

Rational justification is usually understood to require that one acts consistently with one's reasonable beliefs that A ought to be done, even when it is not really (objectively) the case that A ought to be done. Rationality does not demand that one act only on correct beliefs. One need not be irrational in acting on incorrect beliefs, since one may have no sufficient reason to suspect they are incorrect, question and correct them in order to bring them in line with the true balance of reasons. This may suggest that we should adopt the subjective interpretation of NJT: we are justified to obey authority when we have subjective reasons *to believe* this authority helps us improve our conformity to our own reasons.[60]

This interpretation seems supported by a further consideration: the issue at stake is how a right to rule, correlated with a duty to obey is justified. A central feature of duty is that it is wrong if one does not discharge it. Next, it is not only wrong if one does not discharge it – one is blameworthy for failing to do so, as well. If the justification for the duty is in terms of the objectively improved conformity to objective reasons it is likely to bring about, irrespective of one's reasonable beliefs[61] about this prospect, there is no place left for blame in failing to act on the duty. Certain subjective element seems necessary for one to be held responsible for such failures.

However, the favoured by Raz interpretation of NJT seems to be in terms of objectively improved conformity to one's objective reasons alone. Nevertheless, he has recently conceded the need for a subjective element in his account of authority's legitimacy:

> "It seems possible to add a condition for the legitimacy of an authority. Something like a requirement that people over whom it has authority should have reason to find out, and should be able to find out whether it has such authority….Perhaps it should also be a condition of the

[60] This interpretation is suggested by Durning (2003). I find his arguments inconclusive, not least because he does not consider the crucial question whether this interpretation changes the character of Raz's whole conception. Thus the question Durning needs to answer is whether the conception of authority with the suggested by him subjectivist test of legitimacy will remain a conception of practical rather than theoretical authority: purporting to give reasons for action rather than reasons for belief to its subjects.

[61] i.e.what one could reasonably be expected to find out, or to try to find out. The point is not that one could not be held responsible for one's incorrect beliefs. Rather, if one had no reason to suspect that one's beliefs were incorrect, and accordingly did not try to find out, or had no reasonable way to find out that they were incorrect, he cannot be blamed for acting on them.

> authoritative standing of any directive that those subject to it have reason to find out whether it exists and can find out its content" (Raz 2003: 264).

Notice that this concession does not get to the point of admitting that improved conformity to reason is not an objective matter, but rather, a matter of subjective reasons to believe in its likelihood. All it allows is that authority cannot impose an obligation on a subject, even if when followed, it would bring objectively improved conformity to his reasons, if the putative subject did not have any reason to suspect this authority could bring improved conformity to his reasons. This partial subjectivisation of the test of legitimacy tackles the issue of blameworthiness for failing to discharge a duty. It does not make duty fully dependent on one's subjective beliefs, however. Rather, that obedience to authority will bring objectively improved conformity to one's objective reasons is a necessary though not sufficient condition for duty to obey: it is also necessary that one has reason to find out that authority could do this for him, conditional on his obedience. Jointly those two necessary conditions are sufficient for the existence of the duty.

I think Raz is right to resist a fully subjective interpretation of NJT: it will not cohere with his conception of authority as providing objective reasons for action, rather than reasons for belief to its subjects. Raz's conception of authority construes it as practical, not theoretical authority: giving reasons for action rather than reasons for belief.

To demonstrate why the subjective interpretation of NJT would not support Raz's conception of authority as practical, consider the following argument. While it is true that it is not irrational of us to act out of ignorance or on false beliefs, when we have no sufficient reasons to suspect we are ignorant, or have false beliefs, this does not show we have a reason to act on those false beliefs. False beliefs (I wrongly believe I have a reason to obey authority because I wrongly believe it will bring improved conformity to my reasons) cannot generally bootstrap into existence objective reasons for action ("I do indeed have a reason to obey it because it will indeed bring such improvement").[62] Reasons for action and rationality seem to part company here.

[62] Some have, indeed, argued along similar lines. Durning (2003:618-620), for example, suggests it might be justified to induce people to believe they have a duty to obey, even when they have no such duty, since this might bring them closer to acting as they should. Notice, however, that what goes on here is not really "bootstrapping" into existence of a duty to obey – people would have no more duty to obey at the end of the process than in the beginning. The closest to arguing for such a "bootstrapping" comes Edmundson

Notice in this regard, that even less than reasons for action could false beliefs yield a valid duty: recall, that for Raz the right to rule correlates with an obligation to obey authority. Thus if the exercise of authority is justified, it would thereby impose duties of obedience on its subjects. One's beliefs (be they true or false), in terms of which the exercise of authority is justified on the subjective interpretation of NJT, however, could hardly yield (bootstrap into existence) such duties.

Though I believe these considerations in favour of a more objectivist interpretation of NJT to be very important, I should point to a further shortcoming with it. It is not obvious whether the objective interpretation allows to say that authority can itself create duties of obedience. If subjects have to follow authority only whenever it (is more likely to) bring improved conformity to reasons, it is reason, and not authority that obligates. If so, we are back with the rationality paradox of authority: authority does not make difference to how we ought to act. Even if this rather serious problem could be overcome, by working on the qualification "more likely," for example, and showing that it allows for requiring obedience to authority even in cases the latter is wrong, thus bringing in the "practical difference" requirement for practical authorities, there are further problems with this interpretation. I show later in my chapter on the instability of the instrumentally justified strategy to decide to follow authority, that there are serious problems with rationally adopting a strategy to always follow authority on this objectivist understanding of NJT. Thus there seem to be a further support for subjectivising the NJT. This should not come, however, at the expense of neglecting the difficulties with accommodating this thesis within a conception of practical authority, supposed to provide new reasons *for action*, and not just reasons to believe in the correctness of certain reasons for action.

2.3. Cumulative or One-shot Test of Legitimacy?

Neither of these two is the immediately obvious interpretation. Many arguments against the plausibility of Raz's NJT trade on this. There are strong arguments why it should

---

(2002). He argues that authority may justifiably claim to be legitimate (implying a duty to obey it), if by so doing it could change the social meaning of certain actions, and thereby close the gap, opened by the "compliance condition": that people have a duty to comply with authority's commands, only if enough others already comply. The initially incorrect claim helps bring into existence the conditions that activate the duty: enough others begin to obey, thereby obligating the rest to obey as well. This solution is challenged by Lefkowitz (2004).

better be interpreted cumulatively.[63] They have do to mainly with the fact, that unless thus interpreted, it could hardly provide justification for treating authoritative directives as exclusionary reasons, replacing subjects' own, as the Preemption thesis states. It is because authority brings *overall*, and not on each occasion, improved conformity, that its claim to obedience may be justified, even on occasions where acting on its directives is sub-optimal relative to acting on one's own reasons directly. However, this interpretation is in tension with the maximising interpretation of NJT, favoured by Raz (as argued in the next section). The problem is that the maximising logic of instrumental rationality, underlying NJT, may require maximising on each occasion directly and it may not permit the adoption of an indirect strategy of the sort, which favours the cumulative interpretation of NJT. This tension between the cumulative and the maximizing interpretation of NJT is the focus of the discussion in Part three of my thesis. There I also address the difficulties, posed to the maximising interpretation, by the possibility of authority committing great mistakes. Also, I raise there concerns with whether and how could one rationally determine the time-span of beneficially interacting with authority: NJT interpreted cumulatively seems of little help in this regard. It might turn out that rationality requires single, one-shot exchanges with authority: but then the rational benefits of an exchange with authority might be unattainable.

2.4. Maximising or Satisficing?

One very important issue for the plausibility of NJT is whether to interpret it in maximising or in satisficing terms. Apart from the questions raised in the previous section, a further question to address is, whether it is enough that the authority claiming legitimacy over me in a given sphere is better overall than me with respect to issues within that sphere. Or may be it is also necessary that there is no alternative source of directives, that could bring the level of conformity to the reasons within that sphere that apply to me even higher. In short, is it sufficient that authority is good enough (better than me), or it is also necessary that it should be *best* from the available alternatives?[64]

[63] Mian (2002: 105-108)

[64] The subjective/objective and the satisficing/maximising interpretations are discussed by Durning (2003: 602 – 615). This author argues for the subjective and the maximising interpretations, respectively.

This issue will be important when discussing law's and state's claim to supremacy and its compatibility with the autonomy condition in the next part of my thesis.

Raz's interpretation of NJT is not only instrumentalist, but unambiguously maximising as well. One of the main arguments for his Preemption thesis is that treating authority as theoretical, may bring improved conformity to reason. Only treating it as practical authority with pre-emptive, exclusionary force, however, will maximally improve this conformity. If maximally improved conformity is an imperative of rationality, then it follows one is rationally required to obey practical authority (and the Preemption thesis is thus securely established). This claim will be the focus of the third part of my thesis. Let me just point to one possible source of dissatisfaction with the maximising interpretation of NJT. If what rationally justifies my obedience to authority is only that thus my conformity to my own reasons is improved, if there is a better source of directives other than the political authority that makes claims to my obedience, then I cannot be rationally justified to stick to that authority rather than act on the alternative, better directives. Since the logic of justification here is not simply instrumentalist, but maximizing as well, it requires going for the best source of directives, maximizing one's chances of acting in conformity with one's correct balance of reasons. This maximizing rationale may disqualify political authorities as practical authorities for most of their subjects.

Let me just add two more sources of possible concern with this interpretation of NJT. First, it might deepen the problems involved in authority attempting to help its subjects act on their deontological reasons. I have shown[65] why I think deontological reasons may not allow acting for other reasons in meeting their requirements. Even more contestable is that "improved" conformity to them is a desideratum: respecting, not promoting[66] might be all they require. The service of authority would then be in providing the conditions for respecting them – but it is not obvious whether the maximising conception of authority's justification is congruent with such modest service.

Secondly, it is doubtful that one of the central, according to Raz, cases for the justification of political authority – that it could effectively solve collective action problems by helping its subjects coordinate by providing a salient option, demands a

[65] In chapter 1.

[66] Scanlon (1999) distinguishes promoting and respecting value.

maximizing interpretation of NJT. Indeed, all the coordinative function of authority may require is just being "good enough," and not the "best" source of salience.

To brighten up a bit the prospects for the maximising interpretation (after all, I plan to spend a substantial part of my thesis – its third part, dealing with it), let me point to an argument for it. This interpretation seems to go particularly well with Raz's view as to how *reasons can ripen into requirements*.

2.5. Turning "Oughts" into Duties?

In this section, I only very roughly outline the direction of the inquiry to be undertaken in the concluding part of this chapter.

Let me first distinguish between a reason and a requirement. According to Raz, "reason" is what one "ought to do, other things being equal", while "ought" or requirement is equivalent to "has *most* reason to do, all things considered."[67] Requirements are defined in this way on the maximising conception of rationality. On it, reasons legitimate action in terms of advising it. Requirements, or oughts, on the other hand, are mandatory: they do not simply legitimate action – they make it obligatory.

According to the above interpretation of "ought" or "requirements", a maximizing interpretation of the NJT will be favoured. When the conditions of NJT thus interpreted are met, this would close the logical gap between reasons and requirements, both moral and prudential:[68] instead of simply having a reason to do what authority demands, one will be required to do it.

This suggestion, plausible as it may seem, opens the venue for a very difficult inquiry into the further distinction, indeed a gap, between a mere "ought" and a duty or obligation.[69] The role of legitimate authority Raz sees precisely in its "turning 'oughts' into duties."

> "…the difference [an authoritative directive makes] is not in the presence of an additional reason for action, but in the existence of a pre-emptive reason. That is why what is validly required by a

[67] This Razian position of reducing normativity (requirements) to reasons is challenged by Broome (2004) under the name "protantism" (this odd term comes from "ought" as a "pro tanto reason"). Broome agrees that "one ought to F" follows from "one has most reason to F", but doubts that whenever "one ought to F", one also "has a reason to F". Thus the two statements "ought to F" and "has most reason to F" are not equivalent. Normativity and reasons are often divorced (see also Broome (2000)).

[68] On this solution to the gap problem, see Edmundson (2003: 211-214).

[69] In the present context I use these interchangeably.

> legitimate authority, is one's duty, even where previously it was merely something one had sufficient reason to do. Authoritative directives make a difference in their ability to turn 'oughts' into duties." Raz (1986: 60).

What is the difference and does duty add something to mere requirements? Is NJT an adequate test of legitimacy for an authority that would have to turn mere oughts into duties? These are important questions. The complex issues involved demand a detailed analysis, to which I turn in the concluding part of this chapter.

2.6. Deference or Dialogic Model of Authority?[70]

Does NJT as a test of legitimacy for authority favour a dialogic model of authority, involving exchange of reasons, bargaining between the authority and its subjects, or rather supports a deference model (subjects defer to authority's commands)? Raz draws an analogy with an arbitration case to pinpoint the main features of his account of authority.[71] These features, the Preemption thesis being the most important one among them, bring it close to the deference model. The NJT as a test for the legitimacy of authority, on the other hand, has more affinity with the dialogic model, where preemption plays a less significant role. Further, and more importantly, the arbitrator analogy is at odds with NJT: the parties to the dispute have agreed to abide by the arbitrators' decision, irrespective of whether it brings improved conformity to their reasons. (Indeed, the "arbitration model" has been contrasted with a "mediation model," where it is the latter only, which is based on Raz's NJT).[72] The authoritative directive on the arbitrator analogy draws its validity from their agreeing to obey it (arbitrator) rather than from its beneficial character (NJT). This difference is reflected in the different ways subjects' judgements/actions are "pre-empted" by authoritative commands on the two models: only suspended, or rather, altogether abdicated.[73]

---

[70] The distinction "deference" – "dialogic" authority is introduced by Cunliffe and Reeve (1999).
[71] Raz (1986: 41-42)
[72] For this distinction and an argument for the plausibility of the "Arbitration model" as an account of the legitimacy of democratic authority," see Shapiro (2002a: 431-439). The affinity of the arbitration model with democratic authority particularly, is not obvious, however. Thus Cunliffe and Reeve (1999) opt for the Dialogic Model as a model for democratic authority instead. They find the Dialogic model much closer to NJT than to the Deference model (roughly corresponding to Shapiro's Arbitration model).
[73] Cunliffe and Reeve (1999: 461)

Notice, moreover, that the arbitrator analogy gives the strongest support to Raz's Preemption thesis: it is because the parties cannot agree on a solution, that the dispute has been given to an arbitrator to decide. Contesting arbitrator's decision, and not substituting his decision for their own judgement, would defeat the whole point of the arbitration. Hence the plausibility of the Preemption thesis. Preemption, however, is much less plausible on NJT – it seems always permissible, indeed the rational thing to do, to keep an eye on one's ex ante reasons in order to avoid great deviations from them, resulting from following a mistaken authoritative directive.[74]

According to Raz's Preemption thesis, obeying authority (deferring to authority) does not involve surrender of judgement on the part of its subjects – authority requires submission in action, not judgement.[75] So, it will be objected that neither of the two models is faithful to Raz's analysis of authority.

My contention here is that Raz might have downplayed somewhat the importance of surrendering one's judgement in deferring to authority. Notice, first, that acting on CiRs - doing the required by authority action, *because* authority so demands and not because of the evaluative characteristics of this action, does seem to obviate the need for judgement, since there is not much to judge about there. Secondly, not grounding one's action on one's own reasons but replacing them and taking as one's guides for action the authority's ERs, again suggests that one's practical reasoning here is pointless, since one is not to form intention to act as a result of deliberation.[76]

Deferring to authority may thus indeed involve surrendering, even abdicating one's judgement: it is essentially a static, one-way process. Once the dispute is referred to the arbitrator, one is not to argue with him, nor ask for the reasons for his decision in order to evaluate it: one has to submit.

---

[74] These tensions between NJT and the Preemption thesis will be in the focus of my discussion in Part Three of this thesis.

[75] Raz (1986: 39)

[76] According to Shapiro (2002a: 406), Raz neglects this aspect of deliberation: forming an intention to act is the natural final end of deliberation. To say that one can deliberate, but is not to form an intention to act on the result of one's deliberation (when offered an exclusionary reason), does not seem fully consistent. Hart (1982) might then be right in describing authoritative reasons as peremptory (reasons not to deliberate), rather than pre-emptive, or exclusionary (reasons not to act), in addition to being content-independent.

Dialogic Authority, on the contrary, may involve only a temporary suspension of one's judgement, allowing to continuously reevaluate on their merits the authority's claims to obedience: it is a dynamic, two-ways process of exchange of reasons between the authority and its subjects. It allows for accountability of authority, and for taking more fully responsibility for one's life.

NJT seems to give stronger support to the dialogic model of authority, while the Preemption thesis makes more sense on the deference model: these two parts of the Service conception do not seem to be playing in the same team. Whether meeting the NJT condition validates authority's protected reasons with their exclusionary element (as demanded by the Preemption thesis) will be the main issue discussed in the Third part of my thesis. My task here was just to indicate certain tensions within Raz's Service conception of legitimacy.

3. The Coherence of NJT: The Practical Difference Thesis.

The normal way, then, of justifying why subjects should follow a putative authority's directives is that by so doing an improved, and maximally improved, conformity to the reasons, already independently applying to them, is achieved. For Raz political authority is a species of practical authority. What is characteristic of practical authorities is that their directives are meant to make a practical difference to how their subjects should act, and not simply give them reasons to *believe* that the solution authority provides tracks reason better than they themselves would if left alone. This practical, not theoretical, difference is of a special character: authority making such a practical difference is not just to add new reasons, conclusively or not determining what one ought to do. It is to replace the preexistent reasons as well with its own reasons, thus purporting to create categorical obligations, rather than potentially inconclusive reasons.

> Ex: If Alex, on the balance of reasons that apply to him, has to visit a friend, he might have to stay at home instead if an authority with respect to the issues of staying home versus visiting a friend (say, the parent of the child Alex) directs him to do so.

The peculiar thing is that Alex is bound to obey the authority even if the authority (his Dad) is mistaken (as it might well be) on this particular occasion: since Alex's own

reasons have been substituted with authority's, Alex has to act on the latter. The justification for the substitution is that thus Alex will more often act on the correct balance of reasons, then if he acted on his independent reasons directly instead.

This is a simple (and intentionally simplified) picture of how practical authority works. It presupposes that there is a correct balance of reasons, this balance can in principle be known, and it is stable enough to make sense to use it as a benchmark for evaluating the legitimacy of practical authorities.

Political authorities can improve conformity to reason in cases when greater expertise, greater efficiency, stronger will, lack of bias, etc. is what their subjects need. They can also be in a unique position by issuing authoritative directives, to provide solutions to coordination problems, as well as induce cooperation in Prisoners' dilemma (PDs) type of cases. An account of the practical authority of the state has to explain in what precisely does the practical difference authorities presumably make in these cases, consist in.

In the case of expertise-based authority, it is only in a very limited sense that authority makes practical difference to what subjects ought to do. After all, expertise-based authority is the paradigmatic case of theoretical authority – providing reasons for belief rather than action. Nevertheless, Raz maintains that even in cases of expertise, political authorities act on the model of practical authority, thus making practical difference to what subjects should do. I am not sure what motivates this position: certainly it is theoretically more elegant to offer a unified account of authority. For Raz our notion of "authoritative" is best understood on the model of practical authority. This may initially seem plausible.

However, notice the counterintuitive implications of construing expertise-based authority as practical authority. Paradoxically, it is only when this authority is mistaken on a particular occasion, that it could make (if it does indeed) a practical difference to how its subjects should act. In the case of mistake, however, morality or practical reason is not modified, made practical difference to (neither on the spot, nor cumulatively - in a more long-term perspective), but altogether excluded. The exclusion is justified to the extent that the adoption of a strategy to follow authority even in cases when it is, as a matter of fact, mistaken, will eventually bring better conformity to the underlying reasons. It will not, however, bring a modification in morality, or practical reason. This is lucky, because

authority would have a strange effect indeed, were it to be capable of modifying morality through its own mistakes.

The clearest case when authority does presumably make a practical difference without making mistakes concerning the underlying balance of reasons, is when it helps solve coordination problems. This is the central case, according to Raz, also for NJT. It is the core of a justification for political authority in particular. Here presumably it is more difficult to offer[77] an explanation how authority makes a difference to what subjects ought to do by providing reasons to believe (that the underlying balance of reasons is such and such), instead of full-blown reasons for action.

This is, however, a problematic case, precisely because it is very difficult to give a determinate value to whether improved conformity to the subject's own, underlying reasons is or is not brought about. What is brought about may be an improved conformity to the reasons a collective agency (the community as a whole) has to coordinate.[78] Because of the "compliance condition:"[79] one has reason to coordinate only if enough others also coordinate, it is indeterminate whether the reason to coordinate authority provides is dependent on the subject's own reasons (as required by Raz's Dependence thesis) and applies to the subject independently of authority (which is one of the requirements of NJT). [80] If it does not, then whether improved conformity to some reasons is brought about, is irrelevant for the subject's obligation to obey this authority. The new reasons authority claims to give him may not apply to him. [81]

More problematic still is the situation with authority's difference-making role in overcoming Prisoner's Dilemma type situations. Since one does best if all others except him cooperate, and this is true for all, they all defect, thus bringing individually and collectively sub-optimal results. Though it may seem that authority, by inducing compliance, guaranteeing the collectively optimal result of all cooperating, does makes a

[77] Though such challenges were indeed pressed by Leslie Green (1983, 1988) and Donald Regan (1989).
[78] Kutz (2002) has a useful discussion of this problem. He claims that the way to deal with it is by abandoning the individualistic conception of obligation (only within which the "compliance condition" is a problem) and ask instead "What should *we* do to meet [our collective obligation]…." Kutz (2002: 479)
[79] Edmundson's (2002) term for what is better known as the condition of Hart's and Rawls's duty of fair play.
[80] On this ground Waldron (2003) urges a revision of both these moral theses of the Service conception, to bring them close to the "collective" conditions of political obligation in contemporary democratic societies.
[81] I address this issue in more detail in chapter 3, section 3.2.2.1. "Fairness and Efficiency-based Coercion."

practical difference to how subjects ought to act in this type of situation, it is the wrong type of difference and does not really qualify as such.

One central problem is that it is authority's threats of sanctions that make subjects comply with the independent requirements of morality (that they do not defect). Thus authority only provides an additional motivation to comply with the requirements of morality, and not new reasons for action.

Raz's claim, however, is stronger: independently from providing sanctions/additional incentives, authority may help there using its pre-emptive power. There are, however, problems with the compatibility of the coercive element in the directive and its normativity,[82] exacerbated by the fact that authorities seem to rely for making the required here normative type of practical difference precisely on the use of sanctions and incentives. So, normativity may not be conceptually independent from sanctions and incentives after all. Raz indeed recognizes that being a de facto authority is a necessary condition for being a legitimate authority.[83] And de facto authority refers both to the recognition by the subjects of the legitimacy claims of the authority, and to its having actual (non-normative) power over people.[84]

Even if it is conceded that one has to conceptually distinguish the normative power of authority and its use of non-normative, coercive power, serious problems remain. These revolve around the issue whether authority in the "normative power" sense, makes a practical difference to how its subjects should act. Shapiro (2002a: 415) argues that following authority in PDs is either contrary to reason (if reason indeed favours defecting in PDs) or does not make difference to the reasons one anyway has (if reason requires not defecting, in either following indirectly maximizing strategy, or acting to promote the common good).

The conclusion again is that it is difficult to see how authority, on NJT, makes a practical difference to the way its subjects should act. Authority either directs action against reason (in which case it makes a difference only by committing a mistake) or repeats what

[82] My chapter three is an extended discussion of the complex issues involved in this co-existence of coercion and normativity in authoritative directives.

[83] "There is a strong case for holding that no political authority can be legitimate unless it is also a de facto authority." Raz (1986: 56)

[84] Raz (1986: 65)

reason anyway demands, thus not making practical difference to how its subjects should act.

The conclusion to be drawn from the discussion in this section is that showing how legitimate authorities *can in principle* make practical difference, whenever the instrumentalist test of legitimacy specified by NJT is met, is far from being unproblematic. May be instrumental justifications have a tendency to turn practical authorities into theoretical only?[85] Thus, to be able to show how legitimate authority could make such difference to what its subjects should do, one may need to employ a different, non-instrumentalist test of legitimacy. The preceding discussion does not provide conclusive arguments for this conclusion, though such conclusion seems warranted. But I need not go into further arguments to build my case, however: the next section points to even more serious problems with such instrumentalist type of legitimacy tests.

4. NJT and the Moral Duty to Obey Problem

An account of practical authority should be capable of explaining how its subjects could have a duty to obey it – even when it is mistaken. This is again part of the practical difference problem. Authority is legitimate, has a right to rule, correlated with a duty on the part of its subject to obey it whenever subjects' conformity to their own reasons is by obeying authority overall improved. This "Service conception" answer to the legitimacy problem has a strong appeal. On it, one's fundamental independence from authority is established: what ultimately matters, is the individual and his conformity to his reasons, authority is merely a tool for his purposes. This is the humanistic rationale for the instrumentalist justification of authority – only individuals have inherent value.

However, this instrumentalist solution breeds problems. In the preceding section I demonstrated the problems with establishing that practical authorities can make a normative, practical difference to how its subjects should act.

[85] I come back to this point, in discussing the implications of the disjunctive view of normativity and coercion at the end of chapter three.

Next, a no less serious problem is to demonstrate how authority could induce a transition "from ought to duty". I have briefly touched on this in 2.5 above, and it is time to try to give this important issue the treatment it deserves.

For Raz, it is the role of authorities to turn oughts into duties. The discussion on the practical difference thesis above was in terms of making difference to what subjects "ought" to do. May be the reason it was difficult to identify how authority makes a practical difference, is that we looked for the wrong type of difference. The right difference may be that authorities turn mere rational or moral requirements into categorical duties. This suggestion is well worth discussing: after all, duty, not simply requirement is the central notion in the authority-subject relation. The fault with a subject disobeying the legitimate claims of authority is not that he fails to act on a defeasible requirement. Rather, he commits a wrong in breaching an (unconditional) duty.

4.1. The Goal-Independence Condition for Duty

Raz thus distinguishes between "oughts" (requirements) and duties: the former may depend on the agent's goals, while the latter should not:

> "Obligations derive from consideration of values independent of the person's own goals and that is another reason why he is thought of as bound by them despite himself" Raz (1975: 224).

Thus, the distinction between requirements and obligations (duties) relies on the distinction goal dependence/goal independence. Coupled with Raz's position that the role of authorities is in turning oughts into duties, it helps demonstrate one of the problems with Raz's account of legitimate authority.

One objection against Raz's NJT as a test for legitimacy of authority is that it yields at best requirements to obey, but never duties, and as such is inadequate as such a test. It should now be obvious why this objection is warranted. Legitimate authorities turn oughts into duties, the difference between them being that the one is, and the other is not goal-dependent. NJT, however, is insensitive to this distinction. What is especially puzzling, then, is how NJT, the central substantive condition of legitimacy, itself insensitive to the distinction between goal-dependent and other reasons (requiring only that maximally improved conformity to whatever valid reasons there are, goals among

them, is as a matter of fact achieved), could yield goal-independent obligations rather than mere oughts. The grand claim of the critics of Raz's Service conception, and of the instrumental reading of NJT in particular, is that mere instrumental rationality does not yield duties.

One may try to deflect this objection against NJT by saying that NJT (being an instrumentalist test, and thus insensitive to differences in the underlying reasons) need not itself be sensitive to this distinction, in order to yield duties. It works together with the Preemption thesis, and it is the latter, which is responsible for distinguishing goal-dependence from goal-independence.

Thus the work of turning oughts into duties may be done by the Preemption thesis, and not by the NJT. Let me indicate why this suggestion will not help solve the problem.

Firstly, for Raz the Preemption thesis flows naturally from NJT: there is no rationale for the validity of preemption, but for the justification, if there is indeed such, supplied by NJT. It is warranted to treat the directives of authority in the way prescribed by the Preemption thesis, only because this is the single, uniquely available way of reaching the benefits promised by NJT. This functional in essence argument suggests NJT and the Preemption thesis are closely intertwined: without NJT preemption is not rational, without preemption, the benefits of NJT are just promises (cannot be achieved). [86]

Secondly, assuming that the Preemption thesis stands alone, irrespective of its support for/from NJT, it would again not solve the problem of turning oughts into duties. The reason is simple. Being a structural, formal thesis – about the proper way to relate to one's first-order reasons, when addressed with a valid authoritative directive, it is far from clear how it could nevertheless discriminate between goal-dependent and other reasons, since this latter distinction is a substantive issue.

My conclusion is that the Preemption thesis needs a rationale from a substantive thesis that could back such discriminating role: it cannot stand alone. If I am right that NJT cannot help it in discriminating goal-dependence from goal-independence, and oughts from duties, respectively, other substantive thesis should perform this task.

[86] However, recall the discussion in 2.6. above: it indicated the problems with the congruity of these two theses. The concern raised there was that they may be playing in different teams, not in the single "Service conception" team.

To further demonstrate that the Preemption thesis cannot help in distinguishing oughts and duties, nor can it help in explaining how authority helps in turning oughts into duties, consider the fact that the Preemption thesis states only that legitimate authority gives its subjects exclusionary reasons for action. However, for Raz both mere "oughts," supplied by mandatory rules, personal decisions, commitments, etc. and "duties" involve such exclusionary reasons for action. Thus the fact that authority's claim to provide its subjects with such reasons for action, is justified - they indeed are valid, does not show subjects have an obligation to obey rather than just being subject to a conditional (on one's goals) requirement ("ought"):

> "Not all mandatory rules, though, impose obligations. Many of them apply only to persons who pursue certain goals and are binding on them because they help promote these goals…Obligations derive from consideration of values independent of person's own goals and that is another reason why he is thought of as bound by them despite himself." Raz (1977: 224)

Thus, it is goal-independence, not the presence of exclusionary reasons (not the truth of the Preemption thesis claim that legitimate authorities provide such reasons) that elevates duties over oughts. That the justified exercise of authority adds something to these conditional oughts, however, is neither here not there:

> "My exposition of the notion of duty is in terms of its formal character (imposed by mandatory rules) and the kind of reasons on which it is based (not subservient to the agent's own goals)." Raz (1977: 225)

Given the stress Raz puts on the role of goal-independence in distinguishing duties from oughts, I find it especially puzzling why though he claims the role of authority is precisely in turning oughts into duties, he nevertheless does not discuss the goal-independence requirement for obligation in his canonic discussion of the concept of legitimate authority in *The Morality of Freedom.*

On the contrary, in discussing the expertise justification for the exercise of authority, Raz ties it to *dependence on* one's goals specifically. Indeed, when discussing the distinction between being "an authority" and being "in authority over someone," Raz explicitly refers to the dependence on one's goals as a criterion for determining whether an authority *on an issue* is *in authority over* oneself.

> "They [John and Ruth, authorities in Chinese cooking and financial matters, respectively] do not have authority over me because the right way to treat their advice depends on my goals….whether or not there is a complete justification for me to regard their advice as guides to my conduct in the way I regard a binding authoritative directive depends on my other goals." (Raz 1986: 64-65)

It is only when an authority on an issue is at the same time an authority over S on this particular issue, that it imposes *obligation* on S to follow its directives, other things being equal. My contention is that may be it is incorrect to describe the role an expertise-related authority plays in its subjects' practical reasoning, as one imposing obligations of obedience rather than merely providing conditional (on one's goals) "oughts." If so, Raz's analysis of when the claims to obedience political and legal authority necessarily make, are justified, is misleading. If legitimate authority is to turn oughts into duties, some restrictions on the scope of the reasons on which authority operates, will be needed.

My conclusion is that if the justified exercise of authority is to correlate with a *duty* to obey it, NJT should have a more restricted scope. Not all reasons, applying independently to the subjects, just their *goal-independent,* moral (let me add[87]) reasons, should have the potential of ripening into duties, when authority exerts its beneficial powers and brings

[87] One should be mindful here of Raz's position that morality is just a special case of practical reasons. He, furthermore, believes that though, occasionally, one's moral and one's prudential reasons may conflict, the well-being of a morally good person depends on his successfully pursuing goals, advancing an inherent value or the well-being of another: moral and prudential reasons are "intertwined." Raz (1986: 320)

improved conformity to them. It is not clear how close is the family resemblance of this modified NJT with Raz's own NJT, however. The purely instrumentalist character of Raz's legitimacy test, responsible for its humanistic appeal (preserving the independence of individuals in their relations with authority), seems compromised.

4.2. An Instrumentally Justified Categorical Duty?

Apart from this serious problem with Raz's account of the role of authority in turning oughts into duties, I identify a further problem with the categoricity of the duty on an instrumental account of authority's justification. It needs to be addressed independently of the availability of an answer to the preceding concerns. The question, then, is: can a categorical duty of the type we are interested in an account of authority, be created by what is an essentially conditional, instrumental justification? One of the problem is that the duty of obedience is conditioned on the overall success of authority in the relevant sense – improved conformity to reasons.

Let me briefly sketch an argument against an instrumentally justified categorical duty:

1. Authoritative reasons are binding – they give rise to a *duty* on the part of the subject to perform the act they require

2. Duties are categorical, unconditional requirements

3. The bindingness of authoritative directives is established on rational grounds: since one is rationally required to maximize one's conformity to reason, this rationality requirement makes obeying a successful in this regard authority obligatory. Duty is instrumentally justified.

4. Binding authoritative directives, being instrumentally justified, are conditional in at least two senses:

a) firstly, they are dependent on the value of the action they contribute to (if the action lacks value, their contribution, accordingly, also lacks value) – call this the value-condition.
b) secondly, their validity is dependent on whether they are successful in bringing about the valuable action – call this the success-condition.

Conclusion: If binding authoritative directives are instrumentally justified (maximising conformity to reasons), they cannot be categorical requirements (i.e. duties). (1) contradicts the conjunction of (2), (3) and (4).

A further objection to the purportedly categorical character of the instrumentally justified duty of obedience is that when justified in this way, this duty seems overly dependent on the subjects' willingness to conform to reason. If this is true, the duty could only be hypothetical in a further sense than the one of being goal-dependent: it will also be dependent on agent's willingness to maximize conformity to reason. This duty again is not categorical.

A response that: since we do have a rationality obligation to maximise our conformity to reasons, the resulting duty is not conditional (hypothetical) on our willingness to bring such maximised conformity, but categorical instead, quite independent of such willingness, is not available.

First, it is a contentious issue whether we do indeed have this type of rationality obligation: one may at most be rationally required not to act for a defeated reason, and not necessarily to act on the best reason available, thereby improving one's conformity to reasons. This consideration is important in evaluating Raz's position that "what we ought to do" is "what we have most reason to do." Not maximising, but satisficing conception of rational requirement may be the more plausible one.[88]

Secondly, if we do indeed have a duty to maximise our conformity to reason, it would yield a categorical requirement to obey, which is of a wrong type. The objection to disobeying legitimate authority, it was stressed already, is not that it is not *rational of us* to do so. Rather, the objection is that this is *wrong* of us. The objection to disobeying legitimate authority is of a distinctly moral character – the categorical duty of obedience is a moral duty, and the failure to act on it – a moral wrong. This objection again stresses

[88] Notice the implications for moral requirements if the maximising conception of requirement covers moral requirement as well. It would follow that we are *morally required* to do what we have *most moral reason* to do. It goes without saying that this is a demanding conception of moral requirement indeed. Also, the satisficing formulation may fit nicely Raz's views on incommensurability: when one acts on an incommensurable reason, one need not act for the best reason. Rather, for all he knows, he acts for an undefeated reason. Raz (1986: 321-66)

the need to restrict the duty to obey authority to cases when authority serves morally necessary purposes.

In conclusion, let me stress that even if NJT as a test of legitimacy is restricted to apply to moral, goal-independent reasons only, the first objection from the conditional character of the instrumentally justified obligation to obey would still have a bite. Even if the obligation is to obey *morally beneficial* authority, this obligation will be conditional on those authorities' success in actually being beneficial. As was indicated in my first chapter, this conditionality of obligation contradicts the claim authorities necessarily make to provide content-independent, exclusionary duties.

Further, not only this conditionality of obligation in general on the success of the practical authority, but the conditionality of obligation to obey the law on authority serving *moral purposes* specifically, is in stark contradiction with what Raz claims to be an essential feature of political and legal authority in particular. Such authorities necessarily claim comprehensive supremacy over all other normative domains, morality included. This claim is obviously implausible if subjects only have a duty to obey those directives that serve morally commendable purposes. Making obviously implausible claims cannot be a characteristic feature neither of practical authorities nor of political authorities in particular. That is why it is important for Raz to show that authorities' legitimacy need not be restricted to serving well moral purposes alone.

However, if I am right in all the above critiques against the instrumentally justified duty to obey, the most a morally unrestricted instrumental justification can yield is a conditional, hypothetical rational requirement instead. Since this is not the notion of duty of obedience, implied by our common sense notion of legitimate authority, having a right to demand from us such obedience, this shows the prima facie inadequacy of the account. However, there might be strong reasons to prefer that account, reasons that license, urge even, a revision of the common sense notion. Arguments for this might well be available.

5. Conclusion

In this chapter I have provided the ground for the discussion of the success of Raz's Service conception of legitimate authority as an adequate conception of the legitimacy of political authority in general and that of a liberal-democratic political authority in

particular. Different interpretations of Raz's central moral test of legitimacy – the normal justification thesis, were discussed. I have argued that it is best interpreted as providing an exclusively substantive, objectivist, cumulatively instrumentalist legitimacy test, employing a maximizing interpretation of instrumental rationality. I have leveled several critiques against this test of legitimacy. The first was about its compatibility with the practical difference thesis. This thesis, recall, was a central, defining feature of Raz's model of practical authority: when legitimate, it makes practical difference to how its subjects ought to act. However, it is difficult to see how the main applications of NJT in the political domain – that authority help solve coordination problems and overcome PDs, unambiguously allow authority to make such genuine practical difference. I then identify the main problem with this test of legitimacy: it cannot account for the grand, central "practical difference" authority is supposed to make. For Raz, it consists in turning mere "oughts" into "duties." I identify several problems in this respect. First, NJT (and the Preemption thesis) is insensitive both to the goal-dependence/goal-independence and to the moral/prudential reasons distinction. Secondly, there are general problems with providing an instrumentally justified categorical duty: what a morally unrestricted instrumental justification can yield is a conditional, hypothetical rational requirement. NJT then, may indeed be an inadequate test for legitimacy, if authority is indeed to turn oughts into duties.

However, this is not the end of the story yet. After all, a revision of our common sense notion of legitimate authority, and derivatively, of our notion of duty to obey, may be necessary, if the Service conception turns out to be a particularly good conception for the legitimacy of political authority. It certainly has a lot to recommend itself: it is the most fully developed, the most sophisticated reason-based justification for the exercise of political authority. It may also have the potential of being the most adequate reason-based conception of the legitimacy of liberal-democratic authority in particular. It may, further, turn out to be a particularly apt response to the puzzles with rationality, which any adequate conception of authority need to address and attempt to solve. The first set of questions is in the focus of my discussion in the Second Part of my thesis. The rationality advantage of the Service conception is the main topic of the Third Part.

Part Two

The Case of Political Authority

Political authority, a species of the genus practical authority, has its own peculiarities. Raz singles out some of them as central features of political authority, and law more specifically (law being the main authoritative institution of the state, through which the state exercises its main functions). Several of them should be mentioned: the law/political authority necessarily makes a normative claim to legitimacy: it claims to have a right to be obeyed, with a corresponding to this right duty on the part of its subjects to obey it. Operating as it does through the legal system, the state through its law necessarily claims supremacy over all other normative domains.[89] Thirdly, though Raz does not recognise it as a central feature neither of law nor of the modern state,[90] it is undeniable that the law and the state extensively and ordinarily use coercion in exercising their functions. Further, state claims to be (and acts as) the exclusive provider of a wide range of services, for which it exacts non-voluntary payment, etc.

My concern in this part is to explore the mutual compatibility of some of these central features of political authority, as well as to see whether and how they fit within Raz's general conception of practical authority. My conclusion (in the first chapter of this part of the thesis) is that there are serious problems with showing the compatibility of the normative claim authority makes with state's extensive use of coercion. In the second and the third chapters of this part of my thesis, I focus on the normative supremacy claim and discuss its troubled relation with the autonomy condition (it is more important to make some decisions by oneself rather than correctly). Two considerations – the plausibility of the endorsement constraint thesis (even if weakly interpreted), as well as the apparent presence of agent-relative reasons for action, may cast doubt on whether political authority can make a bona fide claim to normative supremacy. If this conclusion seems

[89] "Legal systems are comprehensive…they claim authority to regulate any type of behaviour…They do not acknowledge any limitation of the spheres of behaviour which they claim authority to regulate…They *claim authority* to regulate all forms of behaviour … Legal systems claim to be supreme…" Raz (1990b: 150-151, emphasis in the text)

[90] Raz (1990b: 157-161)

too sweeping, it could at least be argued that a *liberal-democratic* type of political authority specifically, which needs to be particularly sensitive to the restrictions of the autonomy condition - in (either) its endorsement constraint thesis version, and/or the presence of agent-relative reasons for action version - should be even less capable of making such a bona fide claim to normative supremacy. Since making this claim is, on Raz's conception, an essential feature of law, and by implication, of state authority, it is not clear whether and how could Raz's conception cover the case of a liberal-democratic type of political authority. It might, after all, be the case that a central feature of this type of authority is precisely that it refrains from making such a comprehensive claim to supremacy. If so, making this claim cannot be a central feature of the concept, since the case of the liberal-democratic type of political authority falls within the core of the concept of political authority.

A further, more general point (raised already in the second chapter of the first part, but not fully developed there) against the internal coherence of Raz's conception, and not just against its adequacy as an account of our concept of political authority, is further pressed. It is that the general tenor of the Service conception of practical authority – obedience to authority is justified when licensed by morality, seems to go against this normative supremacy claim as a central feature of legal and political authority. Discussing this major issue will constitute the concluding section of this part of my thesis.

Chapter Three
Normativity and coercion

It is a conceptual feature of law, according to Raz, that it necessarily claims legitimate authority[91] to regulate the conduct of its subjects. The modern state, as long as it operates through its legal system, also necessarily claims such legitimate authority.[92] To have legitimate authority is to have the normative power to change another's normative situation: to make a practical difference to how he should act. The practical difference the state claims to make, according to Raz, has a special character. In issuing directives to its subjects the state by way of its law claims to create moral duties of obedience by affecting subjects' practical reasoning through providing them with valid content-independent exclusionary reasons for action: i.e. valid protected reasons for action. What is then distinctive of law, as opposed to mere power, is that it necessarily claims to have that practical effect.

Is this claim to legitimacy, with all that it implies, compatible with modern state's other standing feature – its use of coercion, aimed at guaranteeing compliance with its directives? My aim here is to see whether the use of coercion, a feature of law in modern states, is compatible with what is taken by Raz to be an essential feature of law – that it claims to be a legitimate authority, i.e. it claims the right to impose moral duties of obedience to its subjects.

Raz maintains that it is only law's claim to legitimacy, which is essential for, and revealing of the normative nature of law and political authority more generally. It is this feature, he believes, that allows to distinguish the case of law from that of the threats of the highwayman. Raz, further, does not believe that coercion plays an important role in fulfilling the functions of law, even less that it is an essential feature of law.[93] This is a

[91] "The claim of authority is part of the nature of law" Raz (1994: 215)

[92] "The law – unlike the threats of the highwayman – claims to itself legitimacy. The law presents itself as justified and demands not only the obedience but the allegiance of its subjects." Raz (1979: 158)

[93] To support his position that use of coercion is not an essential feature of law, Raz (1975: 159-160) gives the example of a legal system designed for angels, where there is no need for coercion in order to guarantee the compliance with authority's directives.

position, contested even by some of those, who are much in agreement with Raz's overall project. Andrei Marmor, for example, is ready to accept that use of coercion may not be an essential feature of law, but he insists that it may play much more important role there than either H.L.A. Hart or Joseph Raz have assumed.[94] Notice also, that though Raz does not concede that coercion is a major function of law, or plays a major role in it, he does not deny that law often as a matter of fact uses coercion to enforce compliance with its directives. The crucial for our purpose point is that he does not here find a problem for the peaceful co-existence of law's use of coercion and law's claim to legitimacy.[95]

Meir Dan-Cohen has challenged this picture of law and authority as peacefully combining normativity with coercion. He offered elaborate arguments[96] in support of his view that there is an incongruity between law's essential claim to legitimacy and its use of threats of coercion. He further attributes to this incongruity the contradictory attitudes of respect and defiance we typically have towards law and the state more generally.

In the first part of this chapter I explore in detail Dan-Cohen's intriguing and rather complex arguments to that effect, and evaluate the challenge presented to them by Dori Kimel.[97] My conclusion is that Dan-Cohen's arguments have not been adequately responded. Further, I believe that Dan-Cohen himself has underestimated the strength of his conclusions. His arguments seem to imply that the central case of the normativity of law: its claiming to be a practical authority, and not just to be a source of some in principle dispensable reasons for obedience, is being undermined by law's use of coercive threats. This strong result triggers going back to the discussion of the concept of practical authority as a source of duties to obey. This issue was already introduced and

[94] "Coercion is not an essential feature of law, that is, in some abstract or conceptual sense. However, such a thought experiment can hardly settle the question of whether coercion is a major aspect of law in our society….the coercive aspect of law is actually much more important than Hart and Raz seem to have assumed. The effective ability to impose sanctions is a major service that law provides for its subjects." Marmor (2001: 44.)

[95] "Nor do I doubt that all political authorities must and do resort to extensive use of and reliance on coercive and other threats. Yet it is clear that all legal authorities do much more. They claim to impose duties and confer rights. Courts of Law find offenders and violators guilty or liable for wrongdoing… To threaten is not to impose a duty, nor is it to claim that one does. None of this show, that legal authorities have a right to rule, which implies an obligation to obey. But it reminds us of the familiar fact that they claim such a right, i.e. they are de facto authorities because they claim a right to rule as well as because they succeed in establishing and maintaining their rule" Raz (1986: 26).

[96] Dan-Cohen (1994)

[97] Dori Kimel (2003)

discussed in the preceding chapter: my aim with the discussion in the current chapter is to further clarify the theoretical difficulties involved in this concept.

1. Introduction: The Disjunctive View of Normativity and Coercion

The normativity of law, its claim to create obligations of obedience for its subjects, is incompatible with law's use of coercion, often backing this normative claim. This, in short, is Dan-Cohen's disjunctive view of the normativity of law (or authority more broadly) and law's use of coercive threats. It is directed against the additive view, which defends the compatibility of authoritative norms with the coercive threats, used to enforce compliance with those norms. The latter view was the position of H.L.A. Hart. It is accepted by Raz as well, though he assigns a much less important role to coercion, and as we saw, denies that it is an essential feature of law and state authority. In this part of the text, I will discuss both the arguments Dan-Cohen marshals for his disjunctive view, and the challenge to it, advanced by Dori Kimel. My conclusion is that Dori Kimel has not appreciated the full force of Dan-Cohen's challenge to the additive view. Further, Dan-Cohen himself has restricted the scope of applicability of his arguments to the *non-instrumental elements* in an account of law's authority. Some of his arguments may, I claim, be extended to apply to instrumental accounts as well, thus presenting a stronger challenge to the additive view of normativity and coercion, than he suggests. In any case, more needs to be done than Dori Kimel has done, to restore the credentials of this view.

2. The Intuitive Argument: The Requests - Authoritative Utterances Analogy

Dan-Cohen's first step in establishing the case for the disjunctive view of commands and coercive threats is an appeal to our intuitions. He points out that a request backed by a coercive threat (such as "Please, pass me the salt, or else I'll break your arm!") is an oxymoron, and claims that the same applies to an authoritative command backed by a threat. The analogy works, he believes, since authoritative commands and requests share important features: both are content-independent and source-based, i.e. their validity does not depend on their content, but on the fact that they have been issued by a source with a right to issue valid requests/commands.

Dori Kimel objects. The analogy fails, he claims, because the two do not share an important feature, which precisely is responsible for the intuitive incompatibility between requests and threats: only authoritative utterances but not requests are exclusionary reasons. A request backed by a sanction may be an oxymoron, precisely because requests do not exclude conflicting reasons, while threats are meant to do exactly that. A command backed by a sanction, however, is not an oxymoron, since both commands and threats rely on such an exclusion (if not of the reason, at least of the motivation for acting on it). If there is a residual intuitive discomfort with the particular example of a command backed by a threat, meant to reinforce the intended analogy with requests, Kimel claims, it is to be accounted for by the dis-proportionality between the commanded action and the threatened sanction. Does the threat nullify the normative force of the command it would have had, had it stood alone, in the utterance: Pass the salt, or I won't pass you the pepper? This rhetoric question is meant as a concluding blow on Dan-Cohen's first argument.

Before going into more consequential and important critique of these objections, let me make a short remark. A rhetoric question could be a fitting response to Dori Kimel's rhetoric question above: Is this a case of a command backed by a threat? It rather looks like a conditional offer, and certainly not a coercive one.[98] What thus may account for the intuitive plausibility of Kimel's example, I believe, is that there is nothing contradictory or intuitively implausible in attaching a condition to an offer: indeed, conditional offers are neither unusual nor paradoxical. If the above is really a case of conditional offer rather than a command backed by a threat, and this seems to be at least an equally plausible interpretation of this utterance, it should suffice to dispel the air of obviousness to his last objection.

Now, let me point to a more serious problem with Kimel's counter-argument. Dan-Cohen[99] makes an important, crucial for the success of his arguments, distinction between conflicting and disjunctive reasons, which is, puzzlingly, missing from Kimel's

[98] I need not go here into the discussion of the relation between coercive threats and offers. Suffice it to say, that the offer above does not seem intuitively to be classifiable as a coercive one. This offer would not count as coercive neither on a moralised (the addressee does not have a right to expect that he will be given the salt by the utterer) nor on a non-moralised account of coercion ala Nozick (1969) or Zimmerman (1981). The addressee is not being made worse-off than he would otherwise have been, nor is the utterer either actively or passively preventing him from getting the salt.

[99] Dan-Cohen (1994: 28)

discussion. The disjunctive reasons (of which threats coupled with authoritative commands are an example) are *non-cumulative*: a request for passing the salt is not reinforced by a threat that unless the salt is passed, the addressee's arm will be broken, but is undermined, or cancelled altogether. The non-cumulativity of disjunctive reasons is best seen when such reasons point *in the direction of the same action*: though pointing in the same direction, they *cannot coexist as reasons for the same ac*tion. Consider my reasons for going to a theatre tonight. I could have cumulative reasons for doing so: one is that thus I will please Anna, other is that an interesting play is performed tonight, etc. However, my reasons may be disjunctive, or non-cumulative as well: if one of my reasons to go to the theatre is that I will thus please Anna, my reason to go to the theatre because of the interesting play performed tonight might be incompatible with my first reason, if as it happens, the selfish Anna will only be pleased if I go to the theater in order to please her and not because of the interesting play performed.[100]

Kimel, as we saw, provides an alternative to Dan-Cohen's disjunctive interpretation, explanation of the intuitive incongruity of a request backed by a threat. According to it, recall, the fact that threats are directed against conflicting reasons, that they are meant to undercut their motivating force, is incompatible with request's appeal, which is not normally so directed against conflicting reasons, but is rather meant to provide a reason for the addressee to do as requested. In short, requests do not involve an exclusionary component. Precisely this characteristic of requests is compromised when they are backed with threats, since the latter are necessarily directed against the conflicting

[100] Let me note that Dan-Cohen's example of a non-cumulative reason – that film A is showing and that film B is showing tonight are disjunctive reasons for me to go to the movie tonight, is unpersuasive. Two scenarios of his example are possible, and both do not support his case. First is that there is only one movie on in the theatre tonight: either A or B. Dan-Cohen says that my reason to go because A is showing is non-cumulative with my reason to go because B is showing. But this is a wrong description of the case: I could only have one of these reasons, since it is either A or B that is shown tonight. My valid reason is just one, not two, so there cannot be a case of disjunctive reasons here at all. On the second scenario of a theatre with more than one film showing at the same time tonight, again we need not have a case of non-cumulative reasons. That A and that B are shown tonight might be reasons for me to go to the theatre: I might like both films, and, further, I might be indifferent as to which one of them I see: whichever I can get hold of a ticket for without too much hassle, I will be happy to see. Besides, I might derive a somewhat strange pleasure in knowing that two of my favoured films are shown (pride that my local theatre shows a good taste, satisfaction that my preferences are popular, etc.), even though I could only see one of them at a time. So, the fact that A is showing, and the fact that B is showing can in principle be cumulative reasons, under certain description. I believe that my example of a non-cumulative reason in the text does not lend itself so easily to re-description, threatening its intended persuasive force.

reasons. Kimel's explanation purports to show that command, when backed by threats, need not be similarly compromised, since commands are indeed meant as exclusionary reasons – meant to exclude the reasons against the commanded action. Both commands and threats point in the same direction: excluding or outweighing the reasons against the commanded action. They can, accordingly, co-exist perfectly well. His conclusion is that though there might be in principle cases of disjunctive reasons, a command backed by a threat is not one of them. Kimel generalises (p. 36) this point: a threat attached to any reason with an exclusionary element, makes for a non-disjunctive, cumulative reason for action.

However, Kimel's explanation of the intuitive incompatibility between threats and requests, namely the ostensibly missing from the case of requests exclusionary component, which pits them against threats, does not account for all cases of non-cumulativity of reasons. And a command backed by a threat may indeed be such a case of non-cumulative reason. We could have, I hope to show, non-cumulative reasons that *are* exclusionary at the same time: even though they point in the same direction, they might still be incompatible. Dori Kimel has not offered any argument to challenge such a possibility: his argument was simply that because threats and commands point in the same direction, i.e. they exclude acting/being motivated to act for certain reasons, they are compatible, and in principle cumulative. He does not consider the possibility of having ERs (or reasons with an exclusionary element broadly construed[101]), which though pointing in the same direction (do not do A), are nevertheless incompatible as reasons against A. If, however, non-cumulative reasons with an exclusionary element are indeed possible, it is not clear why one cannot have non-cumulativity and hence incompatibility in the case of a command backed by a threat as well.

Here is an example of ERs that are disjunctive or non-cumulative. Consider a situation, when one's reason not to eat for the pleasure of eating (probably because one has promised not to do so) does not simply conflict with, but is in a disjunctive relationship with the reason not to eat for the dis-pleasure of eating. (I am sorry for this artificial

[101] Strictly speaking, threats do not contain exclusionary element, as explained by Raz's account of exclusionary reasons. Threats are not second-order reasons not to act for certain other reasons, but simply first-order reasons with a variable, though usually considerable weight, directed against acting on some or all reasons one has.

example. In my defense I could only point out that it is rather difficult to offer an example of exclusionary reasons with a distinct enough substantive content, that would allow them to conflict irrespective of the fact that they otherwise 'point in the same direction'. Such examples are difficult to come up with, I believe, because one very rarely gives promises not to do things for a particular reason.[102] Rather, one usually just promises not to do certain things, full stop. Similarly, the commands that are said to create exclusionary reasons also endow them with rather uniform content: one is not to do certain action for any other reason than the fact that one has been so commanded. This makes for a rather "formal" or empty of substantive content reason indeed, that could barely come into disjunctive relationship with another such reason. Be this as it may, my admittedly rather odd example above could be flashed out as a case of conflict between a promise not to eat for pleasure and a command [addressed to an anorexic mazohist, perhaps] not to torture himself by eating for the displeasure of doing so).

Though both reasons here point in the same direction of not eating, by not eating one cannot possibly discharge both: one either does not eat for the pleasure of eating, or does not eat for the displeasure of eating. The two reasons are non-cumulative, or disjunctive: irrespective of the fact that both point in the same direction, the presence of one of them nullifies the force of the other. By this (I admit, rather artificial) example of non-cumulative exclusionary reasons, I believe to have shown that even if Kimel is right to explain the non-cumulativity of reasons provided by requests and threats through the presence of an exclusionary element in the latter only but not in the former, this is not sufficient to show that there cannot be non-cumulativity of reasons provided by commands and threats, which both have such an exclusionary element. We have seen that two exclusionary reasons can also be non-cumulative, or disjunctive. If so, the possibility is open that a command backed by a threat may involve contradiction, if not strictly be an oxymoron. This can be established, however, only on substantive grounds, or at least this is the position, defended in the next string of arguments.

[102] I am not saying that this is impossible. However, it is certainly rather difficult to verify whether one has kept one's promise, even introspectively, due to the always present threat of a posteriori rationalizations. Not to mention that moral hazard problems may also be involved.

3. The Substantive Arguments: Incompatibility of Threats and Authoritative Commands

The next step in Dan-Cohen's argument for the incompatibility of threats and authoritative directives is to show that *purely instrumental* accounts of authority cannot explain the attitude of deference we have towards even mistaken authoritative directives. It is precisely our attitude of deference to authority, which is incompatible with authority's use of threats. This step in Dan-Cohen's argument is missing from Kimel's reconstruction. It is a very important step, however, since Dan-Cohen recognises that were the instrumental accounts of authority adequate, there need be no incompatibility between commands and threats. Often threats are the best means to achieve an otherwise desirable end, and as such their use might be justified on a purely instrumental account of authoritative directives. Thus there need be no case of disjunctive relationship between commands and threats attached to them.

To establish such a case, Dan-Cohen undertakes to demonstrate the need to go beyond such purely instrumental accounts. He believes that the "intermittent," as he calls it, picture of the obligation to obey authority that such accounts of justification yield, is not true to our considered convictions about this obligation. Indeed, the point he makes is that we need a non-instrumental type of reason, that would account for the fact that we often (reasonably believe to be) justified to defer to authority even in cases of its issuing mistaken directives: i.e. in cases where an instrumental account would not explain why we should obey such directives. He next identifies three such justifications for deference: obeying in order to express respect for the law, out of gratitude, or out of identification with the community, whose law it is.

Only at this point is Dan-Cohen ready to advance his central, substantive arguments for the incompatibility of commands with threats: the presence of threats undercuts the rationale for deference. Several arguments are offered to this effect: coercion removes the opportunity to express deference and it undermines subjects' capacity to prove trustworthy. Also, it damages authority's self-image as having the right to command voluntary obedience and deference. Let me concentrate on those arguments seriatim, and see whether the objections to them raised by Dori Kimel hold.

3.1. Threats and *Expressive Reasons*: "Expressive Significance of Deference" Argument

Dan-Cohen argues that threats undermine the expressive potential of showing respect for authority by deferring to its commands. For all both the subject, his fellows, and authority itself know, the act of compliance to its commands might have been motivated by the fear of sanctions, and thus cannot serve its expressive task.

In response, Dori Kimel contends that in criminal law a command, even when not backed by a sanction, does not seem to provide an opportunity to express respect for the law/authority: one has to do what criminal law commands anyway. Next, according to him, it is rather unlikely that one will be inclined to do what the law prohibits but for the threat of sanction. Obeying the law is thus over-determined: it could be explained either by our willingness to act on our moral obligation, or by our normal and quite independent of law disinclination to anyway do what criminal law prohibits rather than by our willingness to express our respect for the law through deference. These two rationale for obedience already by themselves, without the help of threats, undermine the expressive potential of obedience. Thus Dan-Cohen's argument from deference for the incompatibility of threats and orders does not here succeed. In the presence of sanctions, for all the subjects know, their compliance may neither be motivated by fear of sanction nor by respect for the law, but rather by their perceived moral obligation or by their natural disinclination to anyway do the prohibited act. So, the argument from the expressive value of deference may not hold in the context of criminal law. If this contention is warranted, Dan-Cohen's argument will be seriously limited: it is precisely criminal law where orders backed by threats are the rule rather than the exception.[103]

3.1.1. Non-instrumental Reasons for Obedience: Deference *in Order* to Express Respect

[103] Edmundson (1998: 109) makes a similar point in defence of his moralised account of coercion and his position that law is not coercive: since criminal law's proposals are not typically wrongful, and it is precisely criminal law that is considered a paradigm of law's coerciveness, law may not, after all, be coercive. Notice that this argument, it seems, could easily back-fire: probably the alleged non-coerciveness of criminal law on a moralised account of coercion is a strong case against the moralised account of coercion itself.

First, let me note that Dan-Cohen recognises that mixed motives may often explain obedience, and that this will inevitably detract from its expressive potential.[104] But he advances several arguments (explicity, publicity, sufficiency) why it is precisely coercion rather than just the presence of mixed motives, that undermines this potential altogether. Before considering them in detail, let me point out another weakness of the objection. As already hinted, its success depends on not taking seriously the crucial preparatory step Dan-Cohen makes: the thesis of the strict necessity to include a *non-instrumental component* in an adequate account of the normativity of authority and for explaining the deferential attitude to authority we typically have. It is not simply nice if law could also enjoy an otherwise optional expressive quality: it is strictly necessary. If this strictly necessary component - accounting for certain central cases of deference to authority, is undermined by law's use of coercion, then Dan-Cohen's argument may indeed be valid. The arguments "from deference" work only against this background. Thus Kimel's contention about the inapplicability of the deference arguments in the context of criminal law is not warranted.

It is beyond doubt that one has to obey the criminal law, on the assumption that it correctly reflects the underlying reasons that in any case apply to the subjects, and when it helps to improve conformity to them. Thus instrumental justifications can explain why we should obey the law in certain cases. In such case, there need not be a place for an attitude of deference and for expressing respect for the law by obeying it: obedience to law is justified and fully explainable on instrumental grounds. However, the point Dan-Cohen makes, is that when the law does not correctly reflect the underlying reasons, and this is fairly obvious to the subjects, if there is a reason to obey (as he believes there is such a reason), it cannot be an instrumental one. It cannot be instrumental, since the law clearly does not allow subjects better to conform to the reasons that apply to them directly in such a case. It must rather be an expressive one – by obeying the mistaken law one expresses one's respect for the authority that issued it. When law, however, uses coercive threats, the opportunity to express one's attitude of respect is removed. In such a situation, notice, one most likely will not have a perception of a moral obligation to anyway abstain from doing what criminal law prohibits. The situation is perceived as a

[104] Kimel (2003: 36-37).

case of mistake. Further, one might not be disinclined to anyway do the prohibited act. Hence, this might be a clear case, where obedience is called for and explainable on purely "expressive" grounds. If it is indeed the case that the only reason one has in such situations to obey the law, is that by obeying the law one could expresses one's respect for it, and threats do indeed remove that single reason, no reason to obey remains. This seems a legitimate interpretation of Dan-Cohen's conclusion that "coercion weakens, and sometimes removes, one non-instrumental reason for obedience."[105] Law's normativity may thus be undermined in such a case by law's threat of coercion.

Notice that Kimel does admit that the use of threats may be the price (in terms of loss of an expressive potential) to be paid because law is "not purely normative". He nevertheless concludes that "normativity can perhaps be understood, in the legal context at any rate, as requiring obedience *out of* deference, but not necessarily *in order to express* deference" (emphasis in the original).[106] If that were so, Dan-Cohen's argument from the expressive potential of deference would fall short of establishing the case for the disjunctive view of commands and threats.

However, let me point out that the *obligation-out-of-deference* interpretation of normativity is at home (if at all) with an instrumental account of law's authority, but is not sufficient on the non-instrumental picture Dan-Cohen outlines here. Recall that for him, obeying out of deference will not be justified in instrumental terms, when such obedience is not in fact instrumentally beneficial. If obedience even to instrumentally non-beneficial laws is indeed implied by our concept of obligation to obey the law, the explanation, according to him, calls for a non-instrumental account of authority. On the latter, the obedience-out-of-deference requirement is necessarily augmented with a *rationale* for the non-instrumentally justified in such particular cases attitude of deference.[107] The reason to express one's respect for the law is one such fitting rationale

[105] Dan-Cohen (1994: 39).
[106] Kimel (2003: 39).
[107] The instrumentalist theorist, let me note here, also has to offer a rationale for cases of deferential, not-instrumentally justified obedience, unless he takes obedience out of deference as not rational, and not obligatory. He could solve the problem by way of redescription: what is perceived as a case of deferential, non-instrumentally justified obedience, is a case involving obligation to obey, if and only if it is the case that obedience here is demanded as a result of following an optimal, overall instrumentally justified rule. Discussing the success of this strategy of dealing with the problem of obedience in suboptimal in instrumental terms discrete cases, is the topic of the next part of my thesis.

for deferential obedience. If this rationale is indeed undercut by the presence of coercion, which Dori Kimel admits, more arguments need be advanced to rebut here the disjunctive view of threats and authoritative commands.

3.1.2. "Shared Public Meaning of Deference" Argument

Now, the argument rehearsed up to now is supported, I believe, by the spirit of Dan-Cohen's position, even if not explicitly made. Dan-Cohen's explicit arguments for the distinct way in which coercion undermines the expressive potential of obedience are that threats are explicitly and publicly announced, and that they are meant to be sufficient to compel complying behaviour all by themselves. The reason these characteristics of threats single them out as undermining actions' expressive potential to a much greater extent than just the apparent presence of mixed motives for obedience is that the expressive potential of actions is a matter of their shared public meaning. A coercive system does not allow obedience to have "the social symbolic meaning that would make it a suitable medium for expression and communication": even if one does as a matter of fact obey in order to express his respect towards authority rather than out of fear of sanction or out of any other motive, the fact that his act is of this nature will most probably be left unrecognised, and thus its meaning and significance will fail to be communicated. Under conditions of a coercive system, the likely general apprehension towards authority thus will deprive such acts of their intended meaning, rendering them pointless. One important reason for deference is thus obliterated by law's use of coercion. The crucial step in Dan-Cohen's argument is precisely that the expressive acts of deference necessarily have a public meaning, undermined by the public character of threats and their intended use as sufficient to compel compliance all by themselves.

Dori Kimel does not challenge this: as we saw, he indeed agrees with this point, contesting only its significance and, more importantly, its support for the disjunctive view. His conclusion was that though it may be somewhat undermined by the threat of coercion, the normativity of law as interpreted to require "obedience out of deference rather than in order to express deference" need not be altogether obliterated. The disjunctive position is thus not established. However, if the argument above (for the importance of the expressive view of deference for a non-instrumental account of

authority) is correct, there is a stronger support for the disjunctive view than Dori Kimel admits. More arguments, especially against the importance of the non-instrumental, expressive interpretation of deference, need be advanced in order to challenge the disjunctive view on this count.

3.2. Reliance on Citizens' Good Will and Cooperation: "Importance of Being Trusted" Argument

Dan-Cohen's next argument relies on a different interpretation of deference: it may depend on the value of being trusted. As such, it again is undermined by law's use of coercive threats. The argument works in the following way. By treating A's command as a content-independent reason for action, i.e., by deferring to it, the subject B expresses his appreciation for A's reliance on him, and hence of A (authority). By thus proving trustworthy, B boosts his own self-esteem: he is worthy of the trust of the highly valued by him A. If A, however, uses threats of coercion to compel compliance with its commands, it obviously does not trust B, and accordingly, B cannot possibly prove trustworthy, boost his self-esteem, etc. One further non-instrumental reason for compliance is thus removed by law's (authority's) use of coercion.

Dori Kimel uses the same argumentative strategy to rebut this argument as well.[108] Far from supporting the disjunctive view, the arguments from trust only establish that the use of the normative method, if and when not mixed with other methods of intentionally affecting subjects' behaviour, has certain positive side-effects. They would admittedly be withdrawn when this method is coupled with the use of coercion and other non-normative methods of affecting behaviour. But this fact would not show that in such situations of loss, the authority does not create reasons to obey.[109] This conclusion is a replica of his conclusion concerning Dan-Cohen's first argument: normativity need not be understood as requiring obedience *in order to* express deference. Thus even if the reason for obedience to prove trustworthy is removed, others remain, and the most normativity may require is that the way one acts on them is *out of deference*. I need not repeat my response to this replicated argument: the importance of the suggested by Dan-Cohen non-

[108] Kimel (2003: 39-40).
[109] Kimel (2003: 39).

instrumental rationale for obedience is not even considered, let alone responded by this line of reasoning.

Let us, instead, concentrate on a new argument Dori Kimel advances here. As a preliminary, he offers an ingenious elaboration on and development of Dan-Cohen's argument, to bring it close to defending the disjunctive position. The way to strengthen it, he suggests, is to maintain that instead of simply manifesting trust in one's subjects, employing the normative method (expecting that subjects will be capable of acting on content-independent exclusionary reasons) requires of authority a considerable trust in subjects' capacities to evaluate authority's claim on its merits, in order to treat the authoritative directives in such a way (i.e. as content-independent exclusionary reasons for action). By backing one's directives with threats, (or by using any other, non-normative method for affecting behaviour, though may be to a lesser extent), authority manifests "a fundamental lack of trust" in its subjects having such capacity. A necessary precondition of using the normative method is thereby arguably removed. However, Dori Kimel argues that since trust is a matter of degree, and trust and circumspection need not be mutually exclusive, some further support should be given to the position that threats are *incompatible* with the kind of trust required by the normative method.

Such support could be derived from the proposition that threats, and particularly coercive threats, are essentially *offensive*: their use shows a fundamental disrespect for and willingness to act against the (real or perceived) interests of the subjects. Since rational agents could be presumed to prefer using not offensive or less offensive methods for meeting their objectives, authority's (a presumably rational agent) use of threats manifests its belief that the normative method will not succeed alone, and this warrants helping oneself even to a deeply offensive method. Notice also that authority is not simply a rational agent that could be presumed to prefer less offensive methods of affecting behaviour. The claim to obedience authority characteristically makes is explicitly justified in terms of serving the interests of its subjects. The use of coercive threats, however, manifests authority's willingness to go, when necessary, against the (real or perceived) interests of at least some of its subjects. In addition to revealing a deep incongruity of authority's intentions, this use of coercive threats makes the offence of authority's distrust doubly offensive: not only does authority distrust subjects' ability to

be guided by the normative method alone, but it is willing to go against the (real or perceived) interests of at least some of them. Be this as it may, what is important for Dan-Cohen's purpose here is that from this conclusion a support for the disjunctive view can be drawn: the combined use by authority of the normative method and threats thus implies the presence of inconsistent intentions.[110]

Dori Kimel's objection[111] is that this otherwise valid (as boosted by him) argument does not apply to the case of law and institutional authority more generally. He sees two problems with its application. Ascribing determinate intentions or attitudes to abstract entities such as law and authority is anyway risky. Describing the law as displaying a lack of trust, or disrespect for the interests of the subjects, which allegedly accompany the use of threats, as a unified attitude towards the subjects as a whole, is doubly dubious, because these attitudes are at home only in more intimate, inter-personal relations, that do not characterise the relationship subjects – authority.

Now, I will not discuss the first part of the argument: it would lead me into the deep waters of the ontology of collective agency, which I am not prepared to discuss. Suffice it to say that, were this to be a serious concern, it might challenge the ascription of a claim (implying both a belief and a determinate intention) to authoritativeness to law as well, which according to Raz is a central, necessary feature of law, revealing of its nature. And such a challenge would certainly require a more elaborate defense of this point. Nor does this part seem crucial for the success of Kimel's objection: rather than being a necessary condition, it seems a re-enforcing consideration for whatever force there is to the second part of the objection.

Kimel calls the second part of his objection "the argument from the multiplicity of subjects." It raises some difficulties I would like to discuss in more detail. Let me first make explicit the two steps in this second, rather complex argument, that are not explicitly distinguished by Kimel: I believe this comes at the price of obscuring what precisely is at stake here. The problem with the application of the trust argument for the

[110] Dori Kimel is cautious to set this argument for the disfavoured by him disjunctive view in such terms that even its success would fall short of establishing the "logical incongruity" of commands backed by threats: "the tension exposed is merely on terms of the normal conditions for the use of the normative method". Kimel (2003: 42-43).

[111] Kimel (2003: 45).

disjunctive view, then, is that it firstly, relies on ascribing attitudes to authority with regard to its subjects, which only make sense in closer, more intimate inter-personal relations, of which the subject – authority relationship is not an instance. Call this the "intimacy" step. The second step could be labeled the "heterogeneity" objection: attitudes of offence are misplaced, since they cannot be displayed in relationships among plural and possibly heterogeneous subjects. The objection from the multiplicity of subjects in its two parts is formal, not substantive. It concerns the applicability to the case of the authority-subject relationship of the otherwise valid "trust" argument.

However, Kimel adds further, substantive in character objections, which in my opinion come closer to refuting the disjunctive view. These objections, however, are left somewhat underdeveloped. They are a corollary to the multiplicity issue, but go further. Since law and state as practical authorities are to regulate (i.e. guide, control and co-ordinate) the behaviour of multiplicity of subjects with a heterogeneity of interests, capacities, etc., these institutions are to do so "in keeping with exacting standards of efficiency and fairness."[112] These considerations, and not the distrust in subjects' capacities, plausibly explain the general (addressed to all, not simply to the recalcitrant citizens) use by law of coercive threats. Further, instead of displaying distrust, the use of threats may play (and be perceived as having) an assurance function, and thus again be essentially inoffensive. I will have to consider in detail the merits of these substantive objections.

3.2.1. A Formal Objection Rebutted

But let me start with the formal objection. One of the problems with its "intimacy" step is that it is not clear why would one need the presence of an "intimate" relationship to ascribe attitudes of the type Kimel's boosted argument yields. The offence caused by the distrust of citizens' capacities authority displays, may be understandable even when the relationship is not yet intimate or is still in the process of being built. Subjects can quite understandably be offended by an authority which, though claiming a right to be obeyed (implying the *presence of trust* that they *can* so act), i.e. acting on the presumption that the relationship is of a type to warrant ascription of such attitude, nevertheless does not

[112] Kimel (2003: 45-46).

give them the chance of proving themselves trustworthy. Such authority does not give them an opportunity to actually develop or sustain such a relationship. Such an intentionally ambiguous, and blatantly insincere position could certainly be considered offensive even in a non-as-yet-intimate relationship: one could justifiably feel manipulated and abused, when the good of an intimate relationship is denied, while its rhetoric is extensively used against one's (real or perceived) interests.

A related point is that this objection does not make clear why the relationship between subjects and authority, which is presumably sufficient enough to formally ground authority's claim to be obeyed (implying the presence of trust in authority that its subjects can so act), is nevertheless considered not sufficient to ground an attitude of distrust, inducing in its subjects a feeling of being offended, when authority backs its claim with a threat.

A further problem is that it is not clear, whether the objection denies ascribing such attitudes towards authority in its relationship with any of its subjects one by one, or just with respect to its subjects as a whole, in plural. The first interpretation (no trust in authority-single subject relationship) often presented in the extensive "trust" literature,[113] seems supported by Kimel's remark[114] that while the trust objection does not support the disjunctive view in the case of authority, it may apply in the case of personal relations, where the requisite attitudes are at home. But if so, one wonders why Kimel would need the second, "heterogeneity" step of his argument - why should it matter that the subjects

[113] For similar points, see, for example, Claus Offe's distinction between "trusting one's neighbour" and "trusting institutions" in Offe (1999) and Erik Uslaner's between "particularised" and "generalized" trust" , Uslaner (1999) (both articles are included in Warren (1999)). Rom Harre, on the other hand, claims that "the trust relation between a person and an institution is a species of the person-to person relation" Harre (1999: 260). To put the issue in more general terms, it should be noticed that different uses of trust tend to stress the *affective* (and moral) aspect of trust at the expense of the *cognitive* one, or the other way around. Thus it might be argued that for "trusting" political institutions, the most important work is done by the cognitive aspect. Certainly some work is left for the affective aspect as well, but probably not to the extent to warrant attributing such attitudes to the two sides of the relationship, that are most at home in closer, more affective forms of trust than those involved in trusting institutions. Such *cognitivist* position would support Kimel's argument above. I am not sure he would like to go in this direction. To see why, note that if one takes a stronger, purely cognitive position to characterise the trust relation in the case of institutions, however, then one might not be able to resist Russell Hardin's conclusion: trust might not be an appropriate attitude to have towards institutional authorities at all, since "subjects typically may not be in a position to trust [in the cognitive sense] those authorities except by mistaken inference." in Hardin (1999: 23-24). Unfortunately, I cannot do justice to this complex issue here.

[114] Kimel (2003: 51).

are plural, and possibly heterogeneous: since such attitudes cannot be ascribed to the authority-subject relationship anyway, the argument from trust is a non-starter.

Before going into the details and problems with the latter, heterogeneity step, let me raise another concern with the "intimacy step" of the objection. Authority relationship need not necessarily be understood as a distant, indirect relation between law and its subject. One reason this part of the objection does not work, is that it assumes that law issues threats only at the general level, where the general rules for guiding behaviour are issued, but not at the level of administrative directives, which are meant to implement and enforce such general rules. Because at this general level presumably there is no close enough relationship between authority and its subjects, the attitude of offence subjects may seem justified to hold when authority backs its normative demands with threats, is in fact inappropriate. But threats may be present at the lower level of administrative prerogatives, where the general rules are enforced and disobedience to them presumably threatened to be punished. It is at this lower level, that a person might justifiably feel offended by an official who threatens her with dire consequences if she does not do what he orders her to do. Notice that the official, issuing the administrative order, typically uses the normative method: he does not only threaten to punish disobedience, but makes a claim that he has a right to be obeyed – that it is one's duty to obey him voluntarily. If Dan-Cohen's argument about the incompatibility of these two methods (threats and claims of legitimacy) is valid, it is clearly applicable in this setting: the relationship between the official and the subject is sufficiently close to warrant the ascription of the requisite attitudes of distrust and offence.

Indeed, as Kimel does recognise that Dan-Cohen's challenge may have a bite in relationships between individuals, it is surprising he misses the occasion to discuss the case of a relationship between an official and a particular subject to authority. Notice also that this sufficiently close authority relationship will at the same time not even potentially be subject to the second part, the heterogeneity objection, to be considered below: the authority as presented by its official may address a single individual, where the relationship retains its "authoritative" nature. To illustrate the plausibility of this step of reducing the relationship between law in the abstract and its multiple subjects, to a concrete relationship of an official with a single subject, William Edmundson's defense

of a duty not interfere with (rather than obey) an administrative prerogative[115] may help. His claim is that though there might not be even a prima facie duty to obey the law even of a just state, there is a prima facie duty not to interfere with the on-the-spot commands of an officer of a (nearly) just state when they are directly addressed at one and demand one's obedience. I will not further discuss here the implications of this defense for Dan-Cohen's disjunctive thesis: both because Edmundson is clearly on the side of the additive view (coercion and normativity are compatible),[116] and because my ambition is rather limited here. I just want to demonstrate that the authority relationship could be a relationship between individuals. Notice that here even though the official acts in his capacity as an official and not as a private person, he is an individual, so this case could not a priori be excluded from the sphere of relationships between individuals, where Dan-Cohen's arguments presumably do apply. These relationships could in principle be close enough to allow for ascribing attitudes of distrust and offence.

To try to forestall a possible objection to this move, let me again appeal to Edmundson's own response to a similar challenge. He considers the problem of whether redirecting attention from duty to obey law to duty not interfere with administrative prerogatives "exalts submission to the person of another, at the expense of an equal and general subjection to impersonal rules."[117] He believes that his view does not have the consequence of abandoning Rousseau's and Kant's rejection of the subordination to personal and idiosyncratic will of others, because "the administrative prerogatives of a *just state*, and the officials who dispense them, are themselves governed by general rules"[118]. This does not, however, deny the special role the personal relation plays in the duty not to interfere with administrative prerogatives. The claim is precisely that the general rules *by themselves* do not create an obligation to obey – only when "embodied" in the person of the official and directly addressed at a citizen, do these rules, through the

[115] In Edmundson (1998: chapter 3).

[116] The second of the three anarchical fallacies he tackles in his monograph on political authority, Edmundson (1998) is dedicated to disproving the thesis that law is essentially coercive, and to challenging the central place the issue of coercion plays in contemporary political philosophy. Further, in his more recent work (see, for example Edmundson (2002), he undertakes to show the way "behavioural techniques", coercion among them, could be augmented with "semiotic ones" so that political authorities can plausibly claim to transform what is morally optional into morally obligatory requirements.

[117] Edmundson (1998: 58)

[118] Edmundson (1998: 59)

normative claims of the official, create such an obligation. I need not embrace the substance of Edmundson's position (duties only at the concrete level of administrative prerogatives, never at the level of general rules) to make my point here: nothing in principle limits the case of authority relationship to the cases of relationship of authority in the abstract (the law), with its subjects (in the plural).

Drawing on the preceding discussion let me make a more general point here. Dan-Cohen's argument for the disjunctive view can be strengthened by noticing that it is not law's generality per se, that alone explains (though it certainly can explain it) why the use of threats to back laws is deeply offensive: namely because everyone's behaviour and not only that of the recalcitrant falls within the scope of the threats, thus offending the genuinely law-abiding citizens. The disjunctive view is also supported at the level of concrete, quite specific authoritative commands, within concrete authority relations between an official and a citizen. When the official backs its demand for obedience with a threat of sanctions or coercion, as he does when issuing even singular, on-the-spot directives to each citizen separately, he could and often does offend each one separately through the simultaneous use of those two methods for guiding subjects' behaviour.

The second, "heterogeneity of subjects" step of Kimel's objection, concerns not the applicability of the distrust attitudes in an authority relationship, but their appropriateness in a relationship between authority and its subjects in the plural, which typically characterises law. It may be that it is only in the relationship to plural subjects, that attitudes of distrust or disrespect cannot be ascribed to authority since a relationship to plural subjects cannot be sufficiently close and intimate to warrant such an ascription. It will follow, it seems, that though the occasional use of threats by concrete officials somewhat undermines the normativity of their claims to obedience, this does not translate at the general level of law. Law's normativity need not be undermined by the presence of threats of sanction and coercion within its codes.

The assumption underlying this position is not supported by explicit arguments, as if the point that an authority relationship to plural subjects cannot be close enough is too obvious to warrant discussion. But notice that at least in one context, the family, the relationships within it, though among plural subjects, are certainly not devoid of

sufficient intimacy or close-ness to warrant the ascription of such attitudes to one or more of the members towards the rest. The authority-subjects relationship need not be different.

What needs to be added is that because the relationship is between authority and a multiplicity of quite possibly rather heterogeneous subjects, that these attitudes are not applicable. This is what I distinguish as a second step in Dori Kimel's argument proper: not multiplicity alone, but the presence of a quite likely heterogeneity is what defeats the necessary for the ascription of these attitudes "intimacy" of the relationship.

Again, it is not clear why heterogeneity should be sufficient to rule out the ascription of these attitudes. More pertinently, if this argument is correct and heterogeneity of subjects does rule out those attitudes, we can again legitimately ask why this heterogeneity is not sufficient to rule out the ground for authority's claim to be obeyed as well. This claim does imply the presence of trust on the part of authority that its subjects, though multiple and heterogeneous, are sufficiently likely to be able and willing to act for the reason that authority requires so. Making this claim to obedience further presupposes that authority believes (or pretends to believe, where this pretense is not entirely implausible) that with respect to each and all subjects, these subjects will be better off in terms of conforming to their own reasons if they follow the general authoritative directives than if they disobey and follow their own reasons directly instead. Since the claim is general, addressed to all, irrespective of the differences in terms of their respective probable benefits to be drawn from the exchange with authority, if the heterogeneity argument is correct, then it will follow that such claim cannot be *plausibly* made (that is, if it is not to be outrageously implausible, and not just insincere).

In short, if the heterogeneity argument is correct, it may present a challenge to a basic tenet of Raz's theory of law and state authority, on which the boosted by Kimel argument from trust itself depends: that these institutions necessarily claim authority for themselves and that their claims should not be *obviously* implausible. This observation need not present an insurmountable problem for Kimel's objection. One could abandon the position that law and the state necessarily claim authority for themselves (a route some do recommended)[119] which will, incidentally, ask here for reconsidering Kimel's own

[119] Kramer (1999) is but one example.

argument for the incompatibility of trust with threats, which relies on it. One could, alternatively, better try to offer arguments why the objection from the heterogeneity need not affect the plausibility of law's claim to authority. In any case, the implications of this argument deserve a serious consideration.

My objections both to the first step in Kimel's argument, as well as the doubts raised concerning its second step show, I believe, that it falls short of the mark of supporting his position that attitudes of distrust and offence cannot characterise the authority-subject relationship. If so, the otherwise valid "offence" argument for the disjunctive view of threats and authoritative commands may apply to the authority-subjects relationship as well.

3.2.2. Substantive Objections

Even though the *formal*, "applicability" objections do not hold, the *substantive* objections still may challenge the disjunctive view. It is time to consider them in detail. They run as follows. Since law and the state as practical authorities are to regulate (guide, control and co-ordinate) the behaviour of multiplicity of subjects with a heterogeneity of interests, capacities, perceptions of their obligations, etc., these institutions are to do so "in keeping with exacting standards of efficiency and fairness."[120] These considerations, and not the distrust in subjects' capacities, or in their willingness to cooperate with the authority, plausibly explain the general (addressed to all, not simply to the recalcitrant citizens) use by law of coercive threats. The law-abiding citizens need not take offense in that. Secondly, the use of threats may have an assurance function (that the recalcitrant citizens will be made to comply with the directives of authority) and thus again be essentially inoffensive. These two (interrelated) considerations may plausibly account for the fact that law-abiding citizens do not normally feel offended when law and authority back their normative claims to obedience with coercive threats.

3.2.2.1. Fairness and Efficiency-based Coercion

Starting with the first substantive objection. It may be argued that the "fairness and efficiency" defense against the offense argument does not easily fit Raz's account of

[120] Kimel (2003: 45-46)

authority.[121] This is not obvious. Raz himself stresses the importance of efficiency and fairness in coordination matters.

> "Coordination achieves its goal only if the bulk of the relevant population participates. Both efficiency and fairness may be involved. Coordination may fail altogether if it does not enjoy sufficient level of cooperation, and those who cooperate may face greater burdens than would be otherwise required because some people prefer to free-ride." Raz (1990a: 15)

He makes this point specifically when addressing the issue of the relation between legitimacy and the use of coercion. The connection with the assurance function of authority's coercion is straightforward for Raz - coercion on fairness or efficiency grounds is often required for assurance purposes, guaranteeing successful coordination.
The arguments that need be made in support of the claim that concerns with efficiency and fairness do not easily fit Raz' account of authority's legitimacy, are complex, and will lead me away from the main issues discussed here. Suffice it to say, that there seem to be a tension between the "collective" aspect of the "coordination" argument for no-offense, and the individualist character of the authority relation on Raz's account of authority and law's legitimacy. Waldron[122] has urged the need in this regard for some revision both of Raz's Dependence Thesis - the authority should not, according to Waldron, directly rely in its decision on the reasons that independently apply to its individual subjects - and of NJT - the conditions for attributing 'private' authority over a single individual may be different from those attributing public authority over a community as a whole concerning common matters. Raz has discussed this in *The Morality of Freedom*. He notes that his "analysis of authority has concentrated exclusively on a one-to-one relation between an authority and a single person subject to it,"[123] which opens a "gap between the public and the private aspect of authority." [124] It is an issue, then, whether the "patchy" (an issue-by-issue and a person-by-person) way the legitimacy of authority's claim to be obeyed (addressed to a single, separate subject) is

[121] I say a little more on that in chapter 8, section 3.1.2. The arguments a Razian needs to advance in order to defend the compatibility of NJT with justice, need to be developed further.
[122] Waldron discusses the problems involved in Raz's account of authority as solving coordination problems for the society as a whole, in Waldron (2003:59-66).
[123] Raz (1986: 71)
[124] Raz (1986: 72)

established, and, accordingly, the "patchy" character of the obligations, leave sufficient space for the operation of such "large-scale"[125] standards as fairness and overall efficiency, properly at home when subjects in the plural relate to authority. Fairness and collective efficiency standards need not be of central concern in a single subject-authority relationship, where the justification for the exercise of authority is better expertise, better decision-making capacity, or better than person's own resolve to discharge duty. They do play a role here, but only indirectly: thus to the extent subjects have individual fairness-related reasons for actions (based on moral duties towards others), and authority can help them in this respect, it is to that extant and in that regard that this authority is legitimate. Raz's position on this is clear:

> "It is not good enough to say that an authoritative measure is justified because it serves the public interest. If it is binding on individuals it has to be justified by considerations which bind them. Public authority is ultimately based on the moral duty which individuals owe their fellow humans" (Raz 1986: 72).

Thus, fairness and efficiency as general, direct constraints on authority's exercise are in place, when interpersonal relations and issues concerning the community as a whole are at stake: when cooperation for production of public goods, or coordination for the solution of collective action problems, is obligatory for the community as a whole.[126]

However, my contention is that even in such "collective" cases, where standards of efficiency and fairness presumably do apply, a generally law-abiding subject can be offended by an authority that threatens him (on fairness or efficiency grounds) with coercion unless he coordinates with the other members of society while claiming he owes it obedience. For it is conceivable that he might have no obligation vis-à-vis that

[125] "Large-scale" may be an exaggeration, but an illustrative one. May be more aptly the point could be put thus: fairness is a relational standard, difficult to be applied to singular, one-by-one cases. The NJT is a case-by case, even if aggregate (taking a *type* of cases at a time) criterion of legitimacy. It is not to deny that fairness plays a role. However, fairness only enters the picture to the extent that it is a reason, subjects already have: it is not a constraint on the test of legitimacy. If it is the case that the authority can bring improved conformity to a fairness-based reason of the subject, then the condition of NJT could be met. But NJT may be met in many other ways, not having to do with fairness. Any time a reason (moral, prudential, fairness or not fairness-related) a subject has can be better served by acting on a reason authority gives rather than by directly acting on it, the requirements of this thesis are met.

[126] Even this cautious conclusion could be challenged. Christiano (2004: 278-280), for example, argues that NJT is fundamentally at odds with justice: NJT could grant legitimacy even to ferociously unjust states. This is the main reason this author rejects NJT as an adequate account of legitimacy.

authority concerning the issue of coordination[127] - he might be better than the putative authority in identifying cases when coordination is required, and in determining what his contribution should be. And even if authority is generally better than him in these respects, he might still have no obligation to coordinate, and accordingly, no obligation to obey authority that demands such coordination. This is so, since the obligation to coordinate is subject to a "compliance condition": C has an obligation to coordinate if and only if enough others already comply.[128]

A difficult to fill gap opens between the *claim* authorities make that C has this obligation and therefore C should comply with authority's demand and coordinate on X, and the *normative fact* that there is indeed such an obligation, since enough others are already coordinating on X and there is a collective obligation to coordinate on X.[129] Possibly the only clear case when this gap is automatically and unproblematically closed, is the highly unlikely case of obligation, when C's participation is critical for the establishment and the sustenance of such coordination practice, and this is known. Then, if C is likely to refuse to recognise authority and thus fail to discharge his obligation, authority might be justified to use the threat of coercion to ensure that C coordinates. Only in this case fairness and efficiency require C's participation and may warrant the use of behavioural techniques (such as the threat of coercion) for influencing C's actions.

However, it seems possible even in this case for C to justly claim that neither fairness nor efficiency require *his* participation rather than that of, say, D, who still also does not

---

[127] Coordination here is used, following Raz's usage (1989: 1189-1194), in much wider and loser sense than the technical game-theoretic use of "coordination" or that of Lewisian-type convention. It involves issues of coordination proper, cooperation for the production of public goods, or goals "all share or all should have" (an effective legal system being one of them), PDs, etc.

[128] Raz does not, it seems, take the compliance condition to present an insurmountable problem for the coordinative obligation: coordination for him "presupposes that people are not trying to foil one another." People for him ground their actions "on a view as to how others should act or are likely to act." Raz here, I think, is avoiding the problem by not stressing the distinction between basing one's actions on how "others are likely to act" and grounding them on how "others should act." For surely how others should act is very often an open question: the passage from a collective duty to coordinate for achieving a goal all should have, to an individual duty to coordinate, is far from being automatic, and does seem to involve "compliance condition" issues. A solution to this problem is offered by Kutz (2002), who develops an account of an essentially collective duty to coordinate (he calls it "participatory conception" of social obligation) that tries to solve most of those problems. Discussing its success is beyond the scope of my present work.

[129] For a helpful discussion of the gap problem and for an alternative to the coercive method suggestion for solving it with the use of semiothic techniques (changing the social meaning of acts), see Edmundson (2002).

coordinate. This on two grounds. Firstly, different subjects may have different interest in and needs for the "products" of coordination. Then it is conceivable[130] that *fairness* might actually free C from the obligation to sustain the coordinative practice, if D has greater interest and need in X than C, and C's in particular rather than D's participation is not strictly necessary for X. In addition, C may have less capacity than D to contribute to X's "production" or "sustenance" and thus efficiency may require D's participation instead of C's. So, if C has no independent coordination-related obligation in this regard, it is also not clear whether authority can be justified to use coercion against him to back its claim that his contribution is necessary for the successful societal coordination. If C, moreover, is ready to respect authority's unjustified *claim* that he has an obligation to coordinate, and is thus ready to coordinate even without having an independent obligation in this regard, acting in a sense in a supererogatory way (participation in a scheme of coordination may involve costs for C), the threat of coercion that backs authority's claim cannot be seen as innocuous, essentially inoffensive behavioral mechanism for influencing subjects' actions. Law-abiding citizens may justifiably feel offended, when threatened with coercion: their willingness to coordinate, even when knowing that they are not under obligation to do so (either on fairness or on efficiency grounds) is thereby compromised.

3.2.2.2. The Assurance Rationale for Threats

Let us now consider the second, connected to the preceding one, assurance rationale for backing authority's normative claims with threats of coercion: it may have the potential of silencing one's offense in the face of a threatening authority, simultaneously making normative claims to obedience. This suggestion is explicitly discussed and partly dismissed by Dan-Cohen.[131] Dori Kimel does not comment on its success. Let me try to remedy that.

---

[130] I cannot respond here to all objections, this line of thought is certain to trigger. Fairness might not demand C's equal participation, but will demand his proportional to his enjoyment of the goods participation. Even if he is not enjoying a particular good at all, he might be enjoying others provided by the coordinative scheme, etc., and thus might be under an obligation to start and sustain such scheme. It is not here the place to discuss all the good and bad arguments in the vast literature on the duty of fair play. For my purpose in this text it is sufficient to show: the view that fairness warrants the use of coercive threats, which thereby cease to be offensive, is not uncontroversial. It has to be argued for.

[131] Dan-Cohen (1994: fn. 40 at 49).

The explanation for 'no offense' in this case is that the threat of sanctions is perceived as playing purely assurance function. Threats as providing assurance are necessary for meeting the compliance condition of the obligation to obey: one has an obligation only in case enough others coordinate, or obey as well, and this is known. Let me stress from the beginning that threats are not the only way to meet the compliance condition,[132] though they might be the most efficient one, and the most extensively used by political authorities. Nevertheless, they are not strictly necessary in that respect. The use of threats to enhance coordination is not objectionable, if they play pure assurance function.[133] They could play such pure assurance role, however, only in communities composed of entirely law-abiding citizens. The discussion, accordingly, is a thought experiment with little direct relevance for our societies (where law-abiding citizens are the rule, but there are exceptions as well). It is, however, important to test our intuition that there is nothing offensive in the use of threats by authority, when they serve purely assurance functions: community of entirely law-abiding citizens is the clearest case. If it does not support that intuition, nothing would.

Three scenarios of generally law-abiding societies that may use threats as an assurance that all will abide, are distinguished by Dan-Cohen: mutual suspicion, perfect and imperfect mutual trust. In cases of mutual suspicion, threats do not play pure assurance functions. Threats of coercion, and resort to coercion, when necessary, are used to keep from defecting those who are tempted to disobey because they suspect others will disobey as well. The use of threats is offensive here. The case of perfect mutual trust, on the other hand, does not involve the use of threats at all – for assurance or for coercively guaranteeing compliance purposes. It is of no interest here.

The most interesting scenario is that of imperfect mutual trust: all are abiding and know that all are law-abiding, but do not know that all know that. As a consequence, a citizen C may suspect that some may decide not to obey, and based on this wrong belief, may

[132] As already mentioned, Edmundson (2002) discusses the success of alternative, semiotic techniques for meeting this condition.

[133] Bratman (1999), for example, argues that coercion is inimical to true cooperation, and should be excluded from cooperative activity. Shapiro (2002b: 410-11), who offers an account of legal authority in terms of Bratman's jointly intentional activity, stresses that coercion is often precisely the way through which authority guarantees cooperation, and as such cannot be excluded from authority structures. However, Shapiro concedes that coercion is admissible only as a back-up solution – as a corrective of backsliding.

himself refuse to obey. If assured that no one will disobey as a result of suspicion that not all are law-abiding, no one will disobey. The initial assurance will require the use of sanctions, but once all have been successfully assured that all obey, this is no longer the case. Dan-Cohen concedes that in this stage, the use of threats does not conflict with authority's normativity: threats are not offensive – they play pure assurance function, and do not detract from the expressive significance of obedience.

Dan-Cohen need not concede even that much. Notice the inherent instability of the situation of purely assurance-directed threats. If, per impossibile, the use of threats has established a permanent condition of perfect mutual trust – no threats (either coercive or purely assurance-directed) are anymore needed, and thus this situation is of no interest for the argument. Alternatively, in a situation of imperfect mutual trust, where threats successfully assure that no one defects, nor believes that anyone else defects, the need to coerce does not arise. This is so, but only on condition that the threats are credible. The way to ascertain that, however, is not by running a difficult to verify counterfactual test: would authority coerce C to obey, were he to refuse to obey because he mistakenly believes that the others may suspect not all obey, and start disobeying themselves. Even this counterfactual test would, according to Dan-Cohen, impair authority's normativity (because it exhibits willingness to appeal to subjects' inclinations: as we will see, this "appeal to inclinations" argument is a further step in Dan-Cohen's defense of the disjunctive view). Authority's normativity is seriously compromised, when it is realized that the way to ascertain the credibility of threats, is again through an occasional if not permanent resort to difficult to justify, arbitrarily distributed coercion. Paradoxically, in this scenario, the eventual use of coercion for credibility sake will be doubly offensive: it will be entirely arbitrary and undeserved. If all abide, and some have to be coerced so that the threat of coercion is credible, they have to be somehow picked, and since no one deserves to be coerced, there seem to be no non-arbitrary way of doing that. We are in a situation, only to a degree better/different from that of the mutual suspicion scenario: threats, to be credible and play its assurance function, should be at least potentially coercive.

The suggestion that triggered this long discussion was that if threats have pure assurance function, their use by authority will not compromise authority's normativity – they do not

offend the law-abiding citizens to which they inevitably also apply and would not compromise authority's claim to obedience. It seems impossible, however, to find a clear, uncontroversial case of threats, playing a pure assurance function. This conclusion - no threats with pure assurance functions, supports the disjunctive view of normativity and coercion: an authority that resorts to coercive threats while making purely normative claims to obedience, displays an offensive lack of trust in its subjects.

Let me also stress, at the end of this long section, another important point. Dan-Cohen's argument from trust, as boosted by Dori Kimel, is much stronger than Dan-Cohen himself intended it to be. It is not just one non-instrumental reason for deference (the importance of proving trustworthy) that is undermined by law's use of coercive threats. Rather, one normal, if not strictly necessary precondition for the use of the normative method: that it has trust in subjects' capacities to evaluate authority's demands and act on them, is undermined by its use of threats. The manifestation of distrust law's use of coercive threats involves, may offend the subjects and erode their willingness to take authoritative commands as valid content-independent exclusionary reasons for action. If so, it might be the case that in order to be able to make bona fide (or at least not outrageously implausible) claims to legitimacy, law needs to substantially cut on or even give up its use of coercive threats altogether. It is an open question, however, whether what we will be left with, will still be law.

3. 3. Preference for Compliance: "Gift Analogy" Argument

I will not discuss in detail Dan-Cohen's third substantive argument for the disjunctive position: that by using threats law, contrary to what it claims, reveals its preference for subjects' compliance rather than their voluntary obedience. It rests on drawing an analogy of the case of authority making demands for voluntary obedience and a person seriously wanting a gift. The use of threats allegedly compromises both.

I agree with Kimel that "the concept of command does not involve the kind of freedom (whether or not to obey) that the concept of gift involves."[134] Nor does it involve a willingness to assume the risk that disobedience might ensue. So the analogy fails.

[134] Kimel (2003: 48)

Further, though the presence of threats may indeed reduce one's freedom not to obey, even when those threats are coercive, subjects' obedience does not become entirely involuntary, or not intentional: only the presence of physical force or extreme manipulation can make it such.

Let me just note that Dan-Cohen's starting position that law's claim to authoritativeness implies that law necessarily prefers that subjects "voluntarily" (in the strong sense, implied by Dan-Cohen's argument, and not just in the sense of a minimal voluntariness, necessary for the attribution to subjects the acts of obedience) obey it, is shared neither by Raz[135], nor by H.L.A. Hart.[136] The preference for such voluntary obedience does not seem to be considered by them an essential feature of law and authority: even if it were to be compromised by law's use of coercive threats, this need not affect their positions.

Let me also remind that Raz usually distinguishes compliance from conformity (though he does sometimes use them interchangeably[137]) on his instrumental account of legitimacy: *compliance* with law is justified whenever thus an improved *conformity* to

[135]Raz (1979: 30) "…. the law claims that the existence of legal rules is a reason for conforming behaviour. This should *not be confused with the false claim that the law requires conformity motivated by the recognition of the binding force, the validity of the law. It is a truism that the law accepts conformity for other reasons (convenience, prudence, etc.)* …The way to interpret the fact that conformity is required even in the absence of other reasons for it is that the law itself is presented as such a reason. *It does not matter if compliance is motivated by acknowledgement of such a claim*. What matters is the nature of the claim itself." (emphasis added)

[136] Shapiro (2001: 206-7): "Notice that, for Hart, guidance does not contain a direct "motivational" component: it does not require that people are motivated to follow the law simply because the law requires them to do so. It is possible, then, for the rules to guide conduct even though the motive for conformity is the threat of sanctions. Hart's claim that the primary function of the law is the guidance of ordinary citizens does not presuppose that ordinary citizens take the internal point of view toward the law. The law is not particularly interested in the reasons people have for conforming: they may be motivated by their concern for their fellow men or simply by their wish to avoid being punished. The law simply cares that its citizens learn what it is that is expected of them and act accordingly." Hart's notion of guidance then, according to Shapiro, is epistemic rather than motivational. Shapiro disapprovingly recognises that often a much stronger view of guidance has been attributed to Hart: "Perry seems to attribute to Hart a much stronger notion of guidance, that is, where conformity is motivated by the rules themselves rather than by sanctions."

[137] In Raz (1989) and Raz (1990), he explicitly makes the distinction: what ultimately matters is conformity, not compliance to reasons. Compliance's role is secondary, instrumental one: the role of protected reasons (of which exclusionary reasons are element) is to help bring improved conformity to agents' underlying reasons specifically through compliance to the former. The examples given (lucky mistake through miscalculation, complying with a promise, etc.) all point to the role of compliance as distinct from conformity.

underlying reasons is achieved. Law's normative claim to legitimate authority involves a claim that subjects do comply with law's directives. So, even on the verbal side, Dan-Cohen is not right to say that law claims voluntary obedience rather than compliance: compliance is demanded instead. This does not mean that the demand for compliance can be met by acting entirely involuntarily: by being forced or manipulated, *I* do not comply, the act of following orders cannot properly be attributed to *me*. But the short-cut (voluntary action is denied by any form of coercion) Dan-Cohen took here to support his disjunctive view is unsatisfactory.

I suspect the real issue Dan-Cohen aims at lies elsewhere. It is again connected with the question "how, in what ways does the law claim we have to obey it?" and whether this claim is compatible with law's use of coercive threats. It raises, however, not the issue of voluntariness, but rather that of the proper motivation, implied by the demand for compliance with law. Does law necessarily claim we have to obey it only because it so demands of us, and not for any other reason? And if this is so, how to interpret Raz's position that law accepts obedience for any other reason. If law's *claim* to authoritativeness only is an essential feature of law, are there any restrictions on how that claim is to be made: sincerely, plausibly, or it is OK if it is insincere, reckless, speculative, etc.? Can law be characterised by its making of a claim, about which it does not really care: whether it is or how it is met?[138] This is an intriguing question, well deserving extensive discussion. It is not, however, explicitly discussed by Dan-Cohen here.

4. The Formal Argument: "Appeal to Inclinations"

[138] The law's necessary claim to authoritativeness has been discussed by Philip Soper (1989). He asks: can law be characterised by making a claim to authoritativeness (implying a general obligation to obey the law) generally believed to be unjustified (since it is widely known that the existence of such a general obligation to obey the law has been successfully challenged by political theorists)? For further discussion and for an argument that law is not characterised by making a claim to authoritativeness, see Himma (2001). Edmundson (1998; 2002) also addresses the issue "How can an authority sincerely claim to possess a moral power that it has insufficient reason to believe it in fact possesses. And how can such a claim be even approximately true, given the widespread acceptance of the denial of duty by those who have pondered it most carefully?" Edmundson (2004: 221-224). This author agrees it posses a serious problem from an account of authority that conditions its justification on establishing that its claims are warranted. His suggestion in Edmundson (2002) that authority may need to make false claims to tease out meeting the "compliance condition," is challenged by Lefkowitz (2004).

Dan-Cohen offers a last argument for his disjunctive view, which relies on a structural feature of authoritative directives: they are not meant simply as first-order reasons to obey the command but as second-order reasons not to act on (some of) the reasons one has independently of the command. This feature is the *exclusionary* element in authoritative directives. It is precisely this feature of authority – that it claims to provide reasons to its subjects with this special character, which pits its normativity against its use of coercive threats. The character of this feature warrants best describing the argument as structural or formal: this argument does not depend, I believe, on any substantive features of our reasons to defer to authority. Rather, it points to formal features of both threats and commands, which account for their incompatibility. I will discuss it in detail, because it holds the promise of providing a strong argument for the disjunctive view. If successful, it could show how the disjunctive view is applicable even if pure instrumental accounts of authority are indeed satisfactory. Thus, if successful, it will make the case for the disjunctive view of commands and threats stronger than Dan-Cohen himself claims (for him, recall, it applies only to non-instrumental accounts of authority).

The argument rests on two premises:
P1. to be effective, threats rely on the addressees' disinclination to assume the risks of sanctions,
P2. a central feature of commands, explaining their character as a potential source of duties to obey, is that the exclusionary reasons for action, created by them, as a minimum exclude acting on any inclinations: both pro and contra the commanded action.
The conclusion easily follows: threats rely precisely on such inclinations, thus contradicting the exclusionary element in authoritative commands, explaining their character of sources of duties to obey.

The first premise P1 is unproblematic. The interesting premise is P2. Dan-Cohen here assumes[139] that commands provide exclusionary reasons for action. Such reasons are characterised thus: 1) they exclude by kind, not weight; and 2) there is a minimum that an

[139] Following Raz's analysis of authoritative directives, promises, mandatory decisions as involving valid duties to act as these require.

authoritative command must exclude to be such: and this minimum excluded is the recipient's present desires - both pro- and contra the commanded action. Conditions 1 and 2 help distinguish the case of authority with its claim to legitimacy from the gunman's orders (the latter only aims to *control* the actions of another, not to *exercise authority* – have *normative* power, over him). A claim to exercise authority involves a more complex intention than simply aiming at controlling subjects' action, so that they conform with one's orders. The obedience required by authority is not simply conformity but *compliance* with those orders, where action on reasons of a certain kind – inclinations, present desires, etc., is excluded.

Here Dan-Cohen labours hard[140] to show that this Razian rather dogmatically-sounding position is well-supported. He offers it a Kantian, as he described it, interpretation. Both source-less moral duties and source-based (law-derived) duties to obey authority share important features: they are of the same genus, which is revealed in the similarities of the practical reasoning of their respective addressees. In both acting on moral duty and obeying an authority, certain motivations (acting on present desires or inclinations) are ruled out. Rather, the proper motivation is to act out of respect for the moral law/the authority: to be lead by one's rational free will. Whatever the adequate explanation for the proper motivation (and I will not defend the Kantian element in it: acting on duty requires acting *out of one's rational free will*: it is a contentious issue whether it solves more problems than it itself creates[141]), it is clear that the exclusion of desires does seem implied by the language of duty, invoked both by morality and by law in its claims to authoritativeness.

The next step in Dan-Cohen's argument is to notice that the logic of authority's normative appeal prohibits acting on any kind of desires. Since the authoritative commands exclude by *kind*, not by weight, then all present desires and inclinations are to be excluded: not only *against* the commanded action, but desires *for* it as well. This blanket prohibition on acting on inclinations directly contradicts authority's use of

[140] Dan-Cohen (1994: 45-47).

[141] The allusion, of course is, to Robert Paul Wolff's (1970) Kantian-inspired philosophical anarchism. It is precisely because complying with authoritative commands precludes one from being lead in one's actions by one's rational free will (an obligation we are presumed by Wolff to have), that he denies the possibility of any legitimate authority and any obligation to obey the law, as it claims to be obeyed, and because it claims to be obeyed.

threats, since threats are meant to enlist their addressees' present desires (to escape the threatened sanction) as reasons for compliance. Dan-Cohen's conclusion is that

> "an appeal by authority to its subjects' supportive inclinations to the exclusion of hostile ones …something unprincipled and self-contradictory, …is precisely what an authority does when it backs its orders by coercive threats."[142]

Kimel does not object to this Kantian-inspired picture of the duty to obey, and it is contentious whether Raz[143] would accept its description of the exclusionary aspect of the authoritative commands. Even if this Kantian interpretation is not faithful to all the details, it picks a feature of exclusionary reasons: the exclusion of all present desires as appropriate reasons for action. This feature presumably alone suffices to run Dan-Cohen's argument here.

Rather, Kimel's contentions are quite limited, though, he believes, sufficient to challenge this argument. He first objects, that it is not correct to describe threats as "excluding" the hostile inclinations, since threats are not exclusionary reasons - they cannot possibly exclude in the strict sense of the word any reason, desire, etc.

This objection rests, I believe, on a misunderstanding of Dan-Cohen's point here. It is not that threats "exclude" certain reasons or inclinations: it is clear that what excludes the hostile inclinations is the authoritative command rather than the threats. If so, Dan-Cohen's conclusion seems sound. First, it is beyond doubt, that backing a command with a threat licenses acting on one's natural inclination to avoid "a sanction." Second, since the exclusionary reason provided by the order, by excluding "by kind, not weight", blocks acting on any inclinations (both supportive of and hostile to the commanded action), backing the order with a threat, triggering acting on an inclination to avoid a sanction (even though it is an inclination supportive of the commanded action) does contradict the exclusionary element in the order. So, an authority that backs its orders by threats does indeed send a contradictory message.[144] This is so, of course, only if the

[142] Dan-Cohen (1994: 47).

[143] See my discussion of the goal-independence condition for obligations (involving exclusionary reasons for action) in chapter two of this thesis.

[144] There might be a way of showing that there is no incompatibility involved by claiming that the exclusion *by kind* is limited only *within the scope of the application* of the command, and this does not coincide with the domain where the threat applies. As I tried to show in my first chapter, I believe there are

analysis of authoritative orders in terms of exclusionary reasons is correct, but this is a fixed, unchallenged point here.

After dismissing the above misunderstanding, let me also comment on Dori Kimel's second objection. According to him, "the combined message which can be extracted from a conjunction of a command and a threat is that acting on any inclination would be inappropriate, whereas acting on "hostile" inclinations – as well as acting for (some or all) other reasons which conflict with the command – will carry a penalty."[145] His claim is that there is nothing contradictory or unprincipled in this message.

My contention here is that Kimel's characterisation of the combined message of backing commands with threats, describes only part of what is involved, and this part is the less interesting one. What is missing from this description is precisely the mechanism, through which the threat is meant to achieve the required "blocking" of acting on hostile to the command inclinations. The way threats work is precisely by again relying on inclinations: the *relevant* inclination, however, is to avoid a sanction (and not the already anyway excluded hostile to the commanded action inclination to disobey the command, as Kimel suggests). The inclination to avoid sanctions is what both motivates the subjects and explains their "complying" with the command actions. It is acting on this intended as "friendly" inclination, licensed by authority's threat, which contradicts the exclusionary element of the authoritative command's claim to obedience. Recall, exclusionary reasons exclude by kind, and any inclinations, both hostile and friendly to the commanded action, fall within their scope of application.

The correct description of the combined message of a command backed by a threat, then, is that it is indeed inappropriate to act on any inclinations (an implication of the exclusionary element in the command), but authority nevertheless encourages the subjects to act on their inclinations to avoid a sanction. The fact that the additional "encouragement" provided by the threat is conditional on awareness on the part of the authority that some subjects are strongly inclined to act against the commanded course of action, and will disregard the command unless backed by a threat, does nothing to show that the message sent by the authority is not contradictory. For it is surely contradictory

---

problems with determining the scope of the exclusionary reasons. Besides, it is not clear why would the threat not cover the same range of issues as that covered by the command itself.

[145] Kimel (2003: 50).

for authority to encourage (through threats) the subjects to act on what the authority avowedly pronounces (by claiming that its commands provide valid exclusionary reasons for action) as inappropriate ground for action: namely inclinations.

My conclusion is that, as it stands, Dori Kimel's rebuttal to Dan-Cohen's argument from the appeal to inclinations is not successful. My concern, nevertheless, is not with the success of the argument in itself. Rather, I believe that if successful, it establishes much more than Dan-Cohen intended.

5. Conclusion. The Disjunctive View: Unintended Consequences

Dan-Cohen's final argument for his disjunctive view of normativity and coercion builds on Raz's view that the minimum an authoritative command is to exclude, by providing exclusionary reasons for action, is agents' present desires. I believe that the discussion of that argument and its presuppositions is of considerable consequence concerning the success of the disjunctive view of commands and coercive threats in a purely instrumental account of legitimate authority. Dan-Cohen, recall, says that the disjunctive view is not plausible there: even if my doctor coerces me into taking a drug that will cure me, this does not detract from the normativity of his demand that I have to take the drug. The instrumental benefits of his demand save its normative force from being "destroyed" by his use of a coercive threat. It is rather contentious, let me note here, whether the doctor does clothe his demand for my obedience in "duty" language (duty to whom? Me? The doctor? The others?). Rather, he gives me advice, which I should only follow were I to care about my health. The strongest normative consequence of his utterance seems to be "a conditional, hypothetical rational requirement."[146] This also explains why the use of a threat by a doctor to enforce such requirement seems an unlikely scenario, and why the patient's consent is a must in the doctor-patient relationship. But let me leave this aside.

My point is that even if we could speak of a duty here, its source is certainly not in the doctor himself, but in my prudential interest, my duty to my family, etc. In short, the doctor is a theoretical, not a practical authority, just pointing to me in a reliable way where my duty lies (that is, if I have a duty here). As a case of theoretical authority, it is

[146] Recall the conclusion of chapter two: NJT as a test of legitimacy at best yields precisely such type of conditional rational requirements. This again shows the close affinities of NJT with the account of authority taking it as theoretical only.

of little direct relevance for our discussion of political authority, a paradigmatic case of practical authority. The defining mark of practical authority, it was already pointed out at several points in my thesis, is that it is itself the source of duty – not just a channel, with the help of which one learns what one has a duty to do anyway, irrespective of authority. The interesting question, discussed in detail in chapter two, is whether an account of authority that describes it as such a practical authority, can allow for instrumental justification for the duty to obey it. This, recall, is Raz's project of building a coherent instrumental account of practical authority. Now, Dan-Cohen's discussion of his own Kantian interpretation of the exclusionary elements in authoritative directives ends with the conclusion that authority cannot use coercion without undermining its normative claim. This conclusion, he admits, does not apply to the case of morality. This is so, he says, because authority is itself the source of the duty: the exclusionary force of its commands is directly challenged when it itself backs them with coercive threats. Not so with morality: the use of coercion to enforce morality's requirements does not threaten its normative force, because the coercion originates and is exercised "outside" of the source of normativity.

This argument implies, if I understand it correctly,[147] that it is only practical and not theoretical authority that cannot use coercion to enforce its normative demands: since only the former is a source of the duty, not just a conduit for it (as the latter certainly is).

We are now in a position to indicate the stronger (than Dan-Cohen suggests) implications, the unintended consequences of his argument. On this line of reasoning, it follows that to the extent the duty to obey the law is instrumentally justified, it is practical reason (or morality) that is the source of the duty, not authority itself. And in this case the use of coercion is presumably unobjectionable. It is far from clear, however, whether on this scenario we will be speaking of practical authority, as defined by Dan-Cohen, as an originator of duties. If we accept his understanding of practical authority as an originator of duties, *stricto sensu*, it seems to follow that there cannot be an instrumental account of such practical authority. If we want to incorporate the instrumental aspect, we might need

[147] To my relief, I turn out not to be the only reader puzzled by the implications of Dan-Cohen's otherwise extremely sophisticated arguments. Steven Wall (2003: 165) notes that Dan-Cohen's impressive arguments leave the reader wondering what conclusion he wants to draw from them – whether this is an internal critique of instrumentalist justifications of authority, or, instead, urges an unwavering commitment to it.

a mixed (and inelegant – just a minor blemish, - but possibly incoherent as well) picture of authority: combining theoretical and practical aspects. The crucial point here is that the central case of law's and authority's normativity, defined by Dan-Cohen as necessarily being *source-based*, will be delivered by the practical aspect. This central case is compromised (if Dan-Cohen's argument is correct) by law's and authority's use of coercion.

What follows, I think, is that on Dan-Cohen's picture of law's and authority's normativity, it is not just one non-instrumental reason for obedience, that is undermined by the threat of coercion, but rather its central case. The disjunctive view of normativity and coercion will thus be true concerning the central case of authority's and law's normativity. This undeniably strong result prompts a closer look at whether and in what sense does authority need to be a source and originator of duties to qualify as practical authority.

As it was already discussed in the first part of this thesis, according to Raz authority is not, nor can it be the ultimate source of the duty to obey it: morality is what renders, if and when it does, obeying authority morally obligatory. However, this general point does not negate Dan-Cohen's point here, that authority, when legitimate, does itself impose duties of obedience on its subjects. Raz holds that authority is legitimate and does impose a duty of obedience (though not a general one) whenever the claim it makes to impose such a duty is justified. This position[148] is in line with Raz's "practical difference" thesis - that practical authorities make difference to how their subjects should act. Thus for Raz authority *is* a source of duties, even if not ultimately so. If authority is a source of duties, and if the use of coercion undermines authority's normative claims, deemed by Raz essential, necessary features of law and political authority, we face I dilemma. This shows either that the concept of political authority used is flawed - if not only making normative claims, but the use of coercion as well is taken as a central feature of law and the state. Or, alternatively, that this concept excludes authority's use of coercion altogether - in case the use of coercion is just contingently, not essentially associated with law and political authority. In this latter case the concept is not true to basic facts of social reality, which shows that, even if not flawed or incoherent in itself, it still presents

[148] I already advanced some arguments for this position in chapter two of this thesis.

an inadequate conceptualization of the social phenomenon of political authority. The conclusion is that one either needs to abandon the model of practical authority as an adequate model for political authority (because of the states' undeniable use of coercion), or has to provide successful arguments against the disjunctive view, or at least show how this view has a limited application - it concerns only *some* non-instrumental reasons for obedience. Neither has yet been successfully done. In either case, this is a much stronger result than the one intended by Dan-Cohen – to explain our contradictory attitudes of allegiance and defiance toward authority.

Chapter Four

The Normative Supremacy Claim and the Autonomy Condition: A Critique of Ronald Dworkin's Endorsement Constraint Thesis.

Is it true that a goal cannot contribute to a person's well-being unless it is endorsed by that person? And if yes, what is the successful argument for this claim? My aim here is not to address this complex issue generally. Instead, I discuss the case Dworkin makes for "the endorsement constraint" thesis (henceforth ECT): he has offered the most extensive discussion and defense of this thesis up to date.

> "We must not propose, as a fixed social goal for any person, some goal that he himself *could not endorse* as a fixed derivative goal for himself, that is, as a goal he must pursue throughout this life just in virtue of *his* highest-order interest [having as good life as possible]." Dworkin (1983: 29, emphases added)

More precisely, this thesis states that genuine endorsement of one's valuable pursuits, goals, relationships etc., is a *necessary* condition for those pursuits to contribute to one's well-being. Kymlicka's formula well illustrates the point of this thesis:

> "A person's life is improved only if he leads it from the inside and according to his own beliefs about what is worthwhile." Kymlicka (1990: 203 -204)

Discussing this issue is important for my thesis, since I am interested to see whether Raz's account of authority is an adequate account of the liberal-democratic type of political authority. Such adequate account must allow for popular participation in the collective decision-making, as well as for constitutional protection of certain individual rights, among which freedom of conscience is of primary importance. Constitutional protection of certain individual rights could plausibly be supported by concerns with the importance for the value of one's life, of acting on one's own convictions. It is a contentious issue whether and to what extent popular participation in collective decision-making could be established on this ground.[149] The main argument for ECT - for the

[149] The endorsement constraint thesis is a strong argument for popular participation in the collective-decision making only if unanimous agreement is the regulative goal of such decision-making. Only if everyone genuinely agreed with a decision will everyone be considered to act from one's own convictions. Such unanimous agreement is not a reasonable, not even an attractive goal. Majority rule is more realistic, and, I believe, more attractive. It does not to the same degree rely on the endorsement constraint thesis:

relevance of convictions, to be tested is that an option cannot contribute to the value of life of a person, if this option (acting on the reasons its prospective value provides) goes against the convictions of that person.

A way of providing justification for the exercise of state authority is by showing that authority is crucial for enhancing the well-being of its subjects. Such are the justifications for authority, despite many and significant differences, both on Raz's Service conception of legitimate authority, and on Dworkin's account of the duty of the community to treat its members with equal concern and respect. Dworkin's account, when translated in terms of well-being, says that the well-being (the interests) of each individual matters, and matters equally. However, if ECT is correct, an attempt by authority to contribute to the well-being of its subjects may be subverted, whenever following the directives of authority goes against the convictions of its subjects. The authority thus may act in a self-undermining way: by trying to achieve higher levels of well-being for its subjects, it might actually achieve lower ones. Furthermore, Raz's Service conception of authority, on which justification for the exercise of authority is not only that through authority subjects' own well-being is enhanced, but that rather, their conformity to both their well-being-related and their moral reasons is thus enhanced, would also be affected by the success of ECT.

Raz's **autonomy condition:** authority is legitimately exercised only in cases when acting correctly is more important than acting on one's own and according to one's own convictions, is a condition on Raz's NJT, screening out only some of the purported exercises of authority as unjustified. However, if ECT is a plausible thesis about well-being, it might have the much stronger effect than the autonomy condition, of considerably blocking the Normal Justification thesis as a test of legitimacy for state authority. I will concentrate on these latter, wider implications of the endorsement constraint thesis for the autonomy condition and its role in Raz's Service conception of legitimacy, at the end of the current part of my thesis, when the results of this chapter,

---

though it still guarantees an *equal ex ante chance* that each citizen will be allowed to act according to one's own convictions. Clearly, endorsement here is underpinned by egalitarian considerations, though is an important consideration on its own. Notice also that if the plausibility of the endorsement constraint thesis is doubted, it is even more difficult to accept he autonomy thesis, or a privacy principle (persons' pursuits are beyond the scope of the collective decision-making; and are to be determined in accordance with those persons' own convictions), requiring constitutional protection of certain individual rights.

together with those of the next one, will be evaluated. For now, I focus on the endorsement constraint thesis in its normal settings: the theoretically interesting issue whether the well-being of a person could be enhanced by providing him with options, against his own convictions concerning their value for his own well-being. The issue is whether authority would act in a self-defeating way in trying to enhance its subjects' well-being against their convictions. What needs to be explored is ECT: the well-being of individuals cannot be served, if they are made to act against their own convictions concerning what is of value in their lives.

In this chapter I first offer a detailed analysis of the above definition of the ECT: I believe this analytical exercise offers the prospect of illuminating the intuitive appeal of this thesis, as well as allows to delineate the most problematic issues involved. I proceed towards this task against the background of Dworkin's more comprehensive views on personal well-being. Thus, next, I briefly analyze Dworkin's account of well-being in its relation with his views on the types of values we have. My main concern will be to explore the resources of his theory of well-being for meeting the objection "from the impossibility of mistake." This argument has been used to challenge ECT.[150] Though I conclude that this objection has a bite concerning Dworkin's view on well-being and his understanding of endorsement, this argument is not entirely successful on a weaker than Dworkin's understanding of ECT. Even if the argument from the impossibility of mistake is met there, there is the further problem with ruling out certain very sophisticated forms of non-coercive, cultural paternalism. The mark of success for an account of endorsement constraint of the strong type Dworkin suggests, is precisely its successful dismissal of such type of paternalism. However, Dworkin's arguments against cultural paternalism are found insufficient. Nevertheless, even the weak view of endorsement, without ruling out sophisticated forms of paternalism, does support wide restrictions on state authority's legitimate actions. Authority claims the right to command obedience, irrespective of its subjects' convictions, and ostensibly justifies this claim to obedience by reference to its subjects' well-being. If ECT even on the weak interpretation is plausible, state authority's claim would be widely implausible. Restrictions on the scope of that claim, and on the

[150] See Wall (1998: 189 –197).

scope of legitimate authority's actions will be needed, going beyond the requirements of the autonomy thesis. Not only should authority leave its subjects act autonomously when it is more important that they choose for themselves rather than choose correctly. It should also leave them an ample space where they could act on their own convictions. The way to correct misguided convictions concerning central aspects of their life, when this is strictly necessary and feasible, is through carefully designed, sophisticated forms of indirect, cultural paternalism.

1. Endorsement Constraint Thesis (ECT) Analysed

EC can be defined thus:

"Genuine endorsement by one of one's own valuable pursuits, goals, etc., is a necessary condition for those pursuits to contribute to one's well-being";

or, 'endorsement by P of V is a necessary condition for V to contribute to P's W'

Several questions are pertinent in order to unpack this definition:

A. What is meant here by endorsement?
B. What counts as "genuine" endorsement?
C. How to interpret the "valuable" in "valuable pursuits, goals"?
D. What are the implications of endorsement as a necessary condition? The possible interpretations of a "necessary condition;" the problem of the temporal location of the endorsement: is there a conceptual connection between something being a necessary condition for X, and the temporal occurrence of that condition (is it relevant when that condition is satisfied, in order for X to be X), etc.
E. What is the understanding of well-being, used in this definition?

Starting with A, how is "endorsement" understood, particularly in the context of evaluating personal well-being?

I take Sumner's definition of happiness or life-satisfaction, to be a plausible first approximation of the meaning of endorsement (in the context of evaluating well-being):

> "[it]is a positive cognitive/affective response on the part of the subject to (some or all) the conditions or circumstances of her life." Sumner (1996: 156)

The cognitive element of the response here involves *evaluation:* the subject has convictions about what is of value in his life, and the subject takes them seriously precisely because he believes they correctly reveal this value.

Endorsement is often understood much more narrowly: just as a positive affective response to the components of one's life. It is understood in terms of *life-satisfaction*, and not cognition/evaluation: as simply feeling good about one's life, irrespective of whether or not one has reflected and formed more or less stable convictions about one's life and judges it good. This is the position of certain critics[151] of Dworkin's ECT, who have claimed, that to the extent there is something appealing in it, it is to be found in the close link between endorsing the components of one's life and being satisfied with one's life (being happy with it).[152]

Dworkin's own exposition at times lends support to this understanding. According to him, if we are concerned with the value of the pursuits, goals etc. *for* the person, whose well-being is being considered, we need to focus on the "response" that person gives to the "parameters of his ethical situation."[153] This characterisation may seem to favour the life-satisfaction interpretation of EC, but Dworkin's discussion does not unequivocally support it. On the one hand, when discussing the problem of the connection between personal convictions and the good life, he presents the issue in terms of life satisfaction:

> 'It seems preposterous that it could be in someone's interests, even in the critical sense, to lead a life he despises and thinks unworthy. How can that life be good *for* him? We are tempted, then, to say, that ethical value must be subjective after all: having a good life must be a matter of ethical *satisfaction*, which means, that it must be a matter of thinking one's life good." Dworkin (1989a: 76)

On the other hand, however, he immediately adds, that subjective life-satisfaction does not exhaust what determines one's good life. Rather, the good of one's life is determined

[151] See Arneson (1999).

[152] Richard Arneson (1999) assigns the value of subjective life-satisfaction a place among the objective list of values, which any plausible account of well-being needs to incorporate. The place assigned to it is in no way privileged: life-satisfaction does not have priority. At most it plays a significant role only when a danger of falling below some threshold of life-satisfaction is present. This position is at best identical with the strong additive view of endorsement I consider later, and in no way supports the constitutive view Dworkin tries to defend.

[153] See Dworkin (1989a: 80).

by life-satisfaction, *normatively* understood – i.e., as not simply dependent on one's thinking (or " feeling") it so. The follow up of the above quote is thus revealing:

> 'But then the wheel turns again: I cannot think my life good, unless I think that its goodness does *not* depend on my thinking it so (emphasis in the original).' Dworkin (1989a: 76)

It is best to interpret endorsement as normative life-satisfaction. Thus the subjective element, naturally found in the *feeling* of life-satisfaction as 'being content with one's life as the best life for oneself' is supplemented with the normative element of the presence of ethical *convictions*: one's thinking of one's life as the best for oneself does not depend simply on one's thinking/feeling it so.

Thus, the *response* that a person gives to the parameters of his ethical situation, in the first quotation above, is best understood as having two interrelated aspects, roughly corresponding to the two elements – cognition and affection (psychological pro-attitude) in Sumner's definition above. One's response is a matter of "correspondence" of one's pursuits, goals, etc., to the convictions one has about their value. It also involves a psychological element: one is unlikely to respond positively (in the sense of having, or forming a pro-attitude) to a goal, or pursuit, if it goes against the affected person's deep convictions.

Dworkin goes beyond the notion of life-satisfaction simple, (in the subjective psychological,[154] not the normative sense that interests me here). This is manifest in the strict *priority* assigned by him to "ethical integrity", defined thus:

> "the condition one achieves, who is able to live out of the conviction that his life, in its central features, is an appropriate one for him, that no other life he might live would be a plainly better response to the parameters of his ethical situation, rightly judged." Dworkin (1989a: 80)

The introduction of the priority of integrity is, according to Dworkin, supported by a wide-spread intuition that the negative aspect of leading a life against one's convictions cannot be outweighed by the positive features of a substitute life. Obviously, such outweighing could be acceptable on an account of well-being, concentrating on the

[154] Interestingly, the issue of life-satisfaction in the psychological sense is sharply distinguished from the question about the successful life, in Dworkin's revised version of the Tanner lecture, included in Dworkin (2000). See especially Dworkin (2000: 241).

importance of life-satisfaction simple. Thus ethical integrity is not implied by the concept of life-satisfaction simple. The priority of integrity thesis denies that there can be any positive features of a life, which are not endorsed by the person. That is why, when evaluating the well-being of a person, Dworkin is careful to mention both "genuine satisfaction and self-approval"[155] as two distinct indicators, and to stress the priority of the ethical integrity of the person.

Defining endorsement as *life-satisfaction, normatively understood* allows to accommodate both intuitions: the importance of subjective contentment with one's life as well as the priority with respect to this feeling of contentment, of one's normative convictions about the goodness of one's life as the best life for oneself. Endorsement thus defined is an attractive solution to the subjective/objective dilemma in the accounts of well-being: for one's well-being it is jointly necessary that one's life has objectively valuable components, and that those components are endorsed. Whether this solution is correct will be one of the main concerns of this part of my thesis.

B: what is meant by "genuine endorsement"?

The requirement that endorsement be genuine (or authentic) is twofold.[156] On the one hand, it involves concerns with having the endorsement based on an adequate information concerning the components of one's life: this has been called "the information requirement." It has two interpretations: a strong one – the "reality" (ideal information) requirement, and a weaker one – the justification requirement (reasonable belief given the available information). On the other hand, endorsement is genuine only when severe forms of social manipulation (let alone coercion) such as conditioning, indoctrination and forced socialisation are not allowed to determine the assessment of the elements of one's life, and their consequent endorsement. The second aspect is considered the more pertinent, since irrespective of the quality of the information one has, one's endorsement is directly influenced by one's convictions, and when the latter are manipulated, even a perfectly informed endorsement would not be genuine. Dworkin stresses the importance of both aspects. His explicit intention is to disprove a strictly

[155] See Dworkin (1989a: 82).
[156] Here I am following once again Sumner's discussion (1996: 156 – 171).

subjective account of well-being. He, accordingly, insists on the importance of the *reality* condition: the value of a component person P endorses is a necessary condition for its contribution to P's well-being. In this respect Dworkin intends to go beyond the limited information requirement (namely *defeasibility*), accepted in Sumner's and others' more subjectivist accounts of welfare. The second aspect (lack of manipulation) is more explicitly defended, since only if all doubts concerning manipulation of one's convictions are dissipated, would the requirement of the priority of ethical integrity gain plausibility.

On C: how to interpret *valuable* in "valuable pursuits, goals, etc"?
What is the understanding of value, presupposed by Dworkin's account of well-being? This difficult question will be of central concern later in this chapter. The complications I will be considering are due to Dworkin's determination to propose a non-subjectivist account of well-being, while not embracing the full-blown objectivity of what he calls the "impact model" of well-being, premised on a rejected by him, transcendent understanding of value. In this connection it is important to stress, that Dworkin's account is an instance of a widely practiced attempt to develop a "hybrid" theory of well-being: one which takes the middle position between subjectivist and objectivist views on well-being.[157]

Concerning D: In principle, to say of some X, that it is a necessary condition for the existence of Y, one normally means, among other things, that X is to exist prior to, or at least simultaneously with the occurrence of Y. This would be so, were we to be interested in the temporal dimension, and not in the establishment of the connection between the two concepts of endorsement and well-being. The purported connection is that the concept of well-being has as its *constitutive* part the presence of endorsement by the person. The normal temporal sequence, however, is not necessarily dominant in the case of *evaluating* personal well-being. This is so, because the evaluation of well-being may and usually is retrospective. Thus, to say that some component of P's life contributed to the well-being of P, it might be sufficient to establish that P at some point

[157] This trend, according to Sumner (1996: 164), has been already present in Aristotle's discussion of prudential value, possibly in Mill's discussion of the higher and lower pleasures as well, and is now manifest in Joseph Raz's, Ronald Dworkin's and, possibly, James Griffin's accounts of well-being.

in time (before, at the time, or after the occurrence of that component) came to endorse it. There is, admittedly, some bias in favour of pre-, or at-the time-of the occurrence of the component endorsement. This has nothing to do with the evaluation of well-being itself. Rather, it is triggered by doubts of an empirical nature that a subsequent to the occurrence of a component endorsement might not be genuine (in the second sense of "genuine" as "not being manipulated"). Notice, nevertheless, that Dworkin does not in principle exclude the possibility of having a genuine endorsement, occurring after a not entirely voluntary adoption of a goal, career, style of life, etc. This is supported by the example Dwokin gives of a child being manipulated or even outright forced into playing the piano, which does not preclude that child from genuinely endorsing this activity at a later point in his life.

A further problem is whether to interpret "endorsement is a necessary condition" as requiring this endorsement to be active (a strong sense, requiring a *self-conscious involvement* with the purported good), or rather, that it must at least be *passively* present. The latter is a weaker form of endorsement, where only a *willing engagement* is involved.[158] Furthermore, as already shown, endorsement as a kind of response to the circumstances of one's life can be understood as being either simply affective (a pro-attitude, feeling good about it, being happy with it), or as being a cognitive-evaluative one (when one judges the respective components positively as good for oneself), or both. Dworkin himself insists particularly on the cognitive-evaluative aspect of endorsement (the priority of ethical integrity over life-satisfaction[159]), but endorsement as a necessary condition may make more sense, if the primacy of the affective aspect is instead maintained. This has been claimed by some of Dworkin's critics.[160]

It should be made as clear as possible already at this point, that endorsement even in the active sense of self-conscious involvement with a component of one's life is conceptually

[158] On this distinction, see Wall (1998: 191-192).

[159] Or, rather, the constitutive role of the ethical integrity for life-satisfaction, since Dworkin denies that there can be genuine life-satisfaction, where this priority is compromised. The contrast I am pointing at in the body of the text is between life-satisfaction, having integrity as its constitutive part (as it is on Dworkin's favoured *constitutive* view on the relation between personal convictions and the good life), and a different conception of life-satisfaction, where integrity is only "the frosting of the cake," to use Dworkin's own words (as it is on the disfavoured by him *additive* view of this relation). This distinction corresponds roughly to the distinction between life-satisfaction normatively understood, and life-satisfaction simple.

[160] Most notably, by Arneson (1999)

distinct from choice, and autonomy more generally. As the above example with the child-pianist showed, we could have an endorsement of (a component of) one's life not only in the weak but in the strong sense as well, even if no autonomous choice has been involved there. Stronger, endorsement is possible even if that life was thrust upon one.[161]

Endorsement is a necessary condition of well-being on Dworkin's *constitutive view* of the relation between personal ethical convictions and the good life. On the disfavoured by him additive view, even after distinguishing between stronger and weaker interpretation of that view (requiring or not requiring the priority of ethical integrity, respectively),[162] endorsement is not a necessary condition: it is important, either all things considered, or only other things being equal.

Let me stress from the beginning: Dworkin intends to defend the constitutive view, on which endorsement is a necessary condition for good life. However, my claim in this text is that the challenge account of well-being he develops in fact supports a stronger view of endorsement – it already becomes a sufficient condition for well-being, in addition to being necessary one. Dworkin's intention, recall, is to offer a non-subjectivist account of well-being. Whether he is successful in defending the constitutive view, without succumbing to the position that endorsement is a sufficient condition as well, will be one of the main concerns in this chapter. The leading question is: does the challenge view of well-being allow something to be independently valuable, before the issue of its endorsement (only relevant, when questions of *evaluating* personal well-being are discussed) is raised. If this is not the case, it is difficult to maintain a clear-cut distinction between endorsement being sufficient, and its only being a necessary condition for well-being.

---

[161] Mathew Clayton (2002) distinguishes in Dworkin two distinct ways in which identification with (components of) of one's life matter in evaluating how well one's life goes. The first is ethical integrity: a goal, project or relationship contributes to one's life only if one identifies with them. The second goes further: for identification with them is necessary that one has freely identified, chosen them as well. The example with the child pianist demonstrates the difference.

[162] The distinction between the two interpretations of the additive view, as well as their relation to the constitutive view, is introduced by Wilkinson (1996). This author claims that critical paternalism would have been unacceptable on both the constitutive and the strong additive views (because both accept the priority of the ethical integrity of the person, i.e. the priority of her ethical convictions). He shows, however, that Dworkin has failed to establish the case for any view, stronger than the weak additive view, thus failing to rule out critical paternalism.

On *E*: Lastly, what kind of conception of well-being is presupposed by Dworkin's endorsement constraint thesis? I already mentioned that the specific account of well-being Dworkin defends is what he calls the "challenge model:" the good life consists in responding in the right way to the right challenge. It is intended as a non-subjectivist model that allows maintaining a distinction between volitional and critical interests. This distinction is to introduce objective elements in the conception, so that criticism is possible. This conception of well-being is not fully objectivist either: endorsement is a constraint, ruling out purportedly paternalistic, "objective list"[163] understanding of well-being.

The most important question is whether Dworkin's account has the resources of meeting the objections from the impossibility of mistake - a criticism, which has been marshaled against it. This is the task with which I start the argumentative part of my text.

2. Challenge versus Impact Models of Critical Well-being: The Underlying Indexed versus Transcendent Value Distinction

Dworkin's ECT is defended against the background of his challenge model of well-being. ECT has been attacked on the ground that it does not allow for the possibility of mistake.

> "If our endorsement determines the issue [whether or not a particular pursuit, whether coerced or not, added value to our lives], then it is not possible for us to be mistaken. Our decision to endorse or not to endorse would settle the matter. But this would make unintelligible a question we could surely put to ourselves; namely "Did this pursuit add value to my life? When we put this question to ourselves, we are not trying to find out whether we have, in fact, endorsed the pursuit; we are trying to find out whether we should *endorse* it. And for this question to even make sense it must be allowed that we could be mistaken about it." Wall (1998: 194, emphasis in the original)

[163] The term "objective list" accounts of well-being, is owed to Parfit (1986: 499).

This objection naturally translates to the model as well. Can it be met? To understand this critique, it is important to introduce a crucial for the model distinction between volitional and critical interests.

"Volitional interests" are those, which improve well-being whenever one achieves something one wants. They roughly correspond to a desire-satisfaction account of well-being. "Critical interests" are those, which improve one's well-being by achieving things one *should* want to achieve: these are the achievements that would make one's life worse not to want.[164] It is a contentious issue whether these interests correspond to the "objective list" account of well-being.

For Dworkin, failure to satisfy one's volitional interests leads to a decrease in one's volitional well-being. This decrease in volitional well-being need not at the same time upset the ethical value of one's life to oneself. Not so with respect to one's critical interests: when not satisfied, both one's critical well-being, and the value of one's life to oneself are thereby diminished.[165] Both the ethical value and the success of one's life are thus entirely determined according to Dworkin by the success in satisfying one's critical interests. The promise of meeting the objection from the impossibility of mistake can be found here – do critical interests allow one to be mistaken about them and, accordingly, be criticised for this mistake?

Before directly addressing this issue, let me point out that Dworkin's work here is not free from conceptual ambiguities and even confusions, which make its discussion rather difficult.[166] Often Dworkin uses "models of value" and "models of well-being" interchangeably.[167] Thus when the value of a person's life is discussed, Dworkin often slips into talking about that value *for the person.*[168] And this latter is an intuitively

---

[164] See Ronald Dworkin (1989b: 484).

[165] See Dworkin (2000: 242-243).

[166] Richard Arneson in his review of Dworkin (2000), also stresses this problem:

> "As always, Dworkin's prose is elegant and graceful, a sheer pleasure for the reader. This is too bad, I am perversely inclined to think. Dworkin's glittering prose reflects away difficulties that it would be better to absorb and either accommodate or fight. …Dworkin's gift for phrasing enables him to construct word bridges that generate the illusion of spanning problems and objections that rhetoric alone cannot resolve….For an example of what I find troublesome, look at the distinctions he develops between critical and volitional interests and between the challenge and impact models of critical value." Arneson (2002:367)

[167] Dworkin calls the two models (challenge and impact) models of critical well-being, models of ethical value, models of critical value, and uses these terms interchangeably.

[168] See once again Dworkin (1989a: 80).

plausible understanding of personal well-being as distinct from the value of life. Moreover, Dworkin does not take the value of life to be impersonally judged. Rather, for him the value of life is evaluated from the perspective of the person, whose life is concerned; the same is true for personal well-being. On the other hand, in denying that *volitional well-being* contributes to the *value of one's life to oneself*, he clearly intends to distinguish the concepts of well-being and value of a life.

A suggestion that the concept of well-being refers to *volitional* well-being only, while what Dworkin refers to as *critical* well-being is more aptly designated as *the ethical value of a life* is not satisfactory, since it makes an implausible distinction between what is and what is not in one's self-interest. Since Dworkin explicitly identifies well-being with satisfying one' s self-interest,[169] in denying that the satisfaction of one's critical interests contributes to one's well-being, and insisting that it contributes to the value of one's life instead, one denies that it is in one's self interest to satisfy one's critical interests.

In conclusion, the concepts of personal well-being generally and that of a value of life for the person whose life it is, are inter-defined, and the distinction between them unclear.

Back to the distinction volitional/critical interests. It is here that our interest lies: it holds the promise of deflecting the objection from the impossibility of mistake. *Endorsement* plays different role in specifying the two types of interests. For volitional interests endorsement of (desire for) certain pursuits is both a necessary and a sufficient condition for them to contribute to one's volitional well-being. For the critical interests endorsement is only a necessary condition for their contribution to the critical well-being of the person. The interesting question is: does endorsement as a necessary condition allow for the possibility of mistake?

The position "critical interests require endorsement" is far from obvious. The prima facie plausible description of the critical interests is that they are distinguished from the volitional interests precisely on the ground that presence of endorsement (where a degree of *volition*, undoubtedly, does play a part) does not constitute them. They are naturally

[169] In Dworkin (2000: 247) discusses the question about the relation of *self-interest* (in both volitional and the critical sense) to morality. For him the two types of well-being (volitional and critical) are identical with the two types of self-interest (volitional and critical, respectively).

judged from a third-person perspective: that of the reasonable onlooker.[170] Precisely this third-person perspective accounts for how one could be mistaken about what is good for him: whether a pursuit actually advances his critical well-being and the value of his life. In this sense, critical interests could be interpreted as those interests, that a benevolent critic would deem essential for the critical well-being of a person. Such is an "objective list" account of critical interests and well-being. It has the undoubted advantage, that on it the concepts of a mistake and a critique make sense: one could in principle be mistaken about what is in one's critical interest, and be criticised[171] from a third-person perspective – that of the reasonable on-looker. Bringing in the first-person perspective – endorsement, as a necessary requirement for critical interests, nourishes the suspicion that Dworkin's challenge model is not objectivist enough, and does not allow for the possibility of mistake and criticism.

This traditional understanding of critical interests, however, is opposed by Dworkin on the ground that it presupposes an implausible model of critical well-being – namely, the "impact model". On the impact model, 'the impact of a person's life is the difference his life makes to the objective value in the world.'[172] Dworkin opposes to it what he labels a "challenge model" of critical well-being. According to it, the good life consists in responding in the right way to the right challenge: 'a good life has the inherent value of a skilful performance.' This inherent value is determined "*within* lives." Dworkin (1989a: 54, emphasis in the original) The critical interests here are already determined "within" the individual lives, and not in terms of a life's output in the world. But if critical interests are thus determined, does that allow for the possibility of mistake and criticism? The natural standpoint for criticism of mistakes – that of the benevolent critic, or the reasonable onlooker, is abandoned. An advantage of this move, according to Dworkin, is

[170] The objection from the impossibility of mistake Wall introduced precisely after distinguishing two perspectives from which the question whether a pursuit contributed to the well-being of a person can be addressed: that of the reasonable onlooker and that of the person himself. ECT privileges the personal perspectives, but it is only from the reasonable onlooker's perspective that this question makes sense at all. From the first-person perspective it arguably does not make sense to ask whether one should endorse a component of one's life as contributing to its value for his life.

[171] It is not entirely clear what stands behind "critical" in Dworkin's critical interests: whether they are critical for the success of one's life, or whether one could be criticised for not adopting those goals, projects, etc., that would make one's life go best. Both connotations are relevant: the ambiguity here may be lucky, though it hardly furthers the aim of analytical clarity.

[172] See Dworkin (1989a: 55).

that the challenge account of well-being is thus congruent with the favoured by him "indexed account" of value, as opposed to a "transcendent account" of value.

A further, connected feature of Dworkin's challenge model is that endorsement is a necessary condition for well-being. Thus the objective-list account of critical interests (and of critical well-being) is left behind. Crucially for the purposes of Dworkin here, however, the difference between endorsement being a sufficient (as in the case of volitional interests) and its, less ambitiously, being only a *necessary* condition (in critical interests) still allows for the possibility of mistake. Precisely this charge: on Dworkin's challenge model, where endorsement is a necessary condition for an option to contribute to one's well-being, mistakes are not possible, is in the focus of my subsequent discussion.

3. Indexed Value: Weaker and Stronger Interpretations

Several questions should be first addressed. First, what is an indexed ethical value, as distinct from a transcendent one? Secondly, what is appealing in the indexed account? Thirdly, if this account is not without an initial plausibility, does it allow for meeting the objection from the impossibility of mistake? The answer to the charge that the challenge model (and endorsement as a central element in it) cannot meet the impossibility of mistake objection will depend on whether the indexed account of value (which grounds the challenge model) allows for mistakes and criticism in determining where one's critical interests lie.

On the impact model of well-being, ethical, personal value is tied to the impersonal value a life produces (its "product value"[173]). Because of this, ethical value is "transcendent" (outside, irrespective of) with respect to the circumstances of the life, in which it is produced. On the challenge model, on the contrary, the close connection between the ethical and impersonal value is broken. The life's value is not exhausted by the impersonal value, by the product it produces, but has the further value of a skilful performance or a response to the challenge of a life. As a result, this latter value quite naturally is indexed to the particular circumstances of the life at stake. This, Dworkin is intent to show, does not threaten to render the value of a performance - the way the

[173] Dworkin (2000: 258)

challenge of a life is met, subjective. One could arguably still meet the challenge of one's life, as indexed to one's particular circumstances, to a greater or lesser extent. Thus one's life can still have greater or lesser "objective" (but not impersonal!) value, even if this value is "indexed" to the given circumstances.

3.1. The Two Interpretations

Can this last claim be sustained? To answer that question, a closer look at Dworkin's "indexed" account of ethical value is needed. At this level – the level of value, Dworkin again helps himself to endorsement as explaining the difference between his account and the transcendent account of ethical value. On the latter it simply *adds* some value, when present, to the otherwise entirely impersonal value. On the former, it is already a *constitutive* element of value itself.[174] Dworkin also suggests that on his account not simply the presence of endorsement, but the *type* of objective value a pursuit should possess, in order to qualify as a contributor to the well-being of a person, is of a *different kind* than that on the transcendent account.

Distinguishing two interpretations of Dworkin's indexed account may help understand this puzzling "different kind of value." A weaker interpretation - endorsement as necessary but not sufficient condition for ethical value, is suggested by Dworkin's example of the religious life and the life in politics.[175] Even though life in politics may be more valuable per se (meaning objectively so), if this life never becomes endorsed by the person who leads it, its value is if not entirely lost, then at least greatly diminished. The *indexing* of an independent value is done by the presence/lack of endorsement: it has a constitutive role for the indexed value.

This weaker, necessary-condition-interpretation of indexed ethical value does not fit Dworkin's rejection of *cultural* paternalism. Cultural paternalism could be justified on the ground that a challenge of life is more valuable, when chosen from a list of (deliberately pre-selected by someone else), independently valuable options. Dworkin dismisses such possibility, because it misunderstands his challenge model of critical well-

[174] See Dworkin's discussion of the distinction between the additive and the constitutive views on the 'connection between our convictions and the goodness of the lives we lead,' Dworkin (1989a: 50). This discussion introduces the topic of paternalism: what forms, if any, are permissible.
[175] Dworkin (1989a: 79 –80).

being profoundly. Accordingly, we should dismiss the weak understanding of indexed ethical value – it could not ground the challenge model. On that model, according to Dworkin, *there is no correct answer as to what is a valuable challenge, independent of the personal choice of that challenge.* The only 'standards,' determining what can count as a right challenge, are the *appropriate circumstances* for people to decide how to live, and they are not to be determined by consulting some transcendent standards.[176]

This I call the strong interpretation of indexed value. It is, I believe, the interpretation favoured by Dworkin. On it, "divorcing ethical value from ethical choice,"[177] is rejected. The right model of ethical value should fuse value and choice. Here, accordingly, the role assigned to endorsement is much greater: it is *constitutive* of ethical value in a strong sense: it is already both a necessary and a *sufficient condition*.

Notice that on the strong interpretation of indexed value, the term endorsement is misplaced. It suggests that some independent from a personal choice value is made to contribute to the well-being of a person by that person endorsing it. This is obviously not so on the strong interpretation, since the choice of the person is what *makes* the ethical value into what it is. Ethical value is indexed to a *personal choice* of value.

On the strong interpretation, personal choice *determines* ethical value, but may be only partly: Dworkin is again not entirely clear on this point. It is an open question what kind is that part of ethical value, which is not determined by the choice itself. There should be something, beyond personal choice, at least partly determining the ethical value: Dworkin's ambition, recall, is a non-subjectivist ethical value, and full determination of ethical value by personal choice could hardly yield it. The objection from the impossibility of mistake – which is a different way of saying that Dworkin's account of value and well-being is subjectivist, is haunting the indexed account of value on the strong interpretation analysed here.

When *endorsement becomes a sufficient condition* for value to contribute to well-being, this is disastrous for the central distinction of the challenge model – critical and volitional interests. They cannot be distinguished, because for both the presence of endorsement (or

[176] Dworkin does not specify in detail what are the standards determining those appropriate circumstances, except that he maintains that some of the circumstances are to be normative, and among those concerns with justice are to play primary role.

[177] Dworkin (1989a: 85)

here better *choice*) of a pursuit is already a *sufficient* condition for that pursuit contribution to one's well-being. So, if Dworkin favours the strong interpretation of indexed ethical value, his challenge model becomes a model of *volitional* rather than critical well-being, where the objection from the impossibility of mistake cannot be met. The argument is simple: (1) one cannot be genuinely mistaken about one's volitional interests (one can, of course, be mistaken about the factual issues of what are the means for serving one's volitional interests, but not about the standards, establishing what counts as one's volitional interests in the first place), and (2) nothing distinguishes systematically volitional from critical interests, since both have endorsement as sufficient condition, (3: conclusion) one cannot be genuinely mistaken about one's critical interests as well.

This is not the end of the story yet. If it is possible to specify some *external constraints* on the personal choice (endorsement) of value, this objection could be met. Such at least prima facie possibility I see in Dworkin's specification of the right circumstances for a choice of a valuable challenge.[178] After all, if one's life goes best when one responds in the right way to the right challenge, this implies that one could respond in the wrong way to the wrong challenge, i.e. be mistaken. Is this response to the objection more than rhetorical?

3.2. The Appeal of Indexed Value

Let us first see what is appealing in Dworkin's account of indexed value, and whether its appeal justifies the risk involved in blurring the distinction between critical and volitional interests.

The argument in favour of Dworkin's account draws extensively on the shared conviction that

> "there is no such thing as the single good life for everyone, that ethical standards are in some way *indexed* to culture and ability, and resource, and other aspects of one's circumstances, so that the best life for a person in one situation may be very different from the best life for someone else in another" Dworkin (1989a: 48 emphasis in original).

[178] I will extensively discuss this issue in the next section.

The indexing of ethical value is plausible, because it allegedly accords with the fact that different types of life can be of value: it makes little sense to say that there is only one type of good life, which is good for different people in different cultural and personal circumstances.

My claim is that all that this argumentative move establishes is that people can have different lives, which lives nevertheless can in principle all be good for those people. It does not say why these differences should necessarily be explained by thoroughly indexing value to the circumstances of those lives. An equally plausible explanation would be that value should not be confined to one single type, but rather, that there are many ways of good life, realising diverse values. Such values might still in a sense be transcendent. A pluralism of transcendent ethical value could, instead, be the answer.

Dworkin's further argument for the indexed account relies on the shared convictions that not only can we have different kinds of good life, but that the good life for us is somehow *indexed* to the particular circumstances of our life. This is Dworkin's response to the anticipated objection above that there can be many good ways of life, realising plural transcendent values. If there is a case in favour of the indexed account, it should trade on those latter convictions. The idea is not implausible on its face: certainly, one can expect that the good life of the stockbroker will be different from the good life of the wildlife preservationist, and the difference will be explainable at least in part in terms of the different circumstances of their respective lives.

Dworkin's point, however, goes beyond that shared intuition. It is stronger, and revealing of his position. Ethical value for him is a function of a "personal response to the full particularity of situation", where there is "no single right response, just a set of these."[179] More radically, ethical value is a matter of "*making* ethical value from nothing."[180] (emphasis added) Notice again the distinction (which Dworkin does not find necessary to notice, even less to discuss) between these last two suggestions. The former makes value dependent on a *response* (where there should be some standards for a minimally satisfactory response, if not for an optimal one), while on the latter, value is a matter of *creatio ex nihilo*, where no comparable standards are in principle available. One wonders

[179] Dworkin (1989a: 66)
[180] Dworkin (1989a: 64)

what is left from the initial "right response to the right challenge" formulation of the challenge model: no single right response, just a set of these and no right circumstances – rather creation from nothing. Recall that it was the "right response to the right challenge" formulation, which was holding the promise of dissipating the objection from the impossibility of mistake. May be its role was only rhetorical, after all. With the disappearance of the right response and the right circumstances of the challenge, however, the possibility of mistake itself dissipates, and thus the objection from the impossibility of mistake cannot be met.

Neither of these ideas, I believe, is implied by our shared conviction that value is indexed to the particular circumstances of our lives. One could imagine a world in which there is a single right response to one's given circumstances of life, which response could endow this life with value. There is no a priori reason to suppose that our world is not such.

Nevertheless, a further, important consideration may reveal why the suggestions above are initially plausible. They might be implications of the correct observation that the value of a life should not entirely depend on its circumstances. Since one's circumstances are mainly a matter of luck, if the value of one's life depended entirely on its circumstances, the value of one's life would also be a matter of luck. And this is a morally unacceptable position. The role of personal choice of value here may be thought to be in making the value of life not entirely dependent on luck in one's circumstances. Since we do not believe that ethical value should be contingent in such a way, we might be willing to embrace the personal choice solution. The second, more radical idea of creating ethical value from nothing obviously goes beyond the idea of indexing value *to the circumstances* of one's life. The indexing of value here is directly *to radical personal choice*, and is not mediated in any way by the circumstances of the choosing person's life. Not so in the case of the more modest idea of having a set of acceptable responses to the circumstances of one's life.

It seems to me that Dworkin's account of indexed ethical value encompasses elements of both of the discussed "implications" of our shared conviction that the good life for us is somehow indexed to our circumstances. His indexed ethical value involves a double relativisation of the good life for a person – to its circumstances and to personal choice.

The main reason one may be resistant to such a view, taking ethical value as fully indexed (in the sense of this double relativisation to the circumstances of life *and* to personal choice of value), is that it threatens to render the idea of normativity of the good incomprehensible. It would in effect deny that values are in principle general. The idea of being guided by entirely private rules (reduced to one's own convictions), determined by entirely private good (fully indexed or identical with one's circumstances), makes little sense.

A connected consequence of this double relativisation is that the possibility of being mistaken about what is good, or of value for one's life, also becomes incomprehensible: it is denied that there are some general standards, deviations from which constitute a mistake. How can one be mistaken about what is good, if this good is good only and exclusively *for oneself*, and is being judged as good once again *by oneself* alone? Dworkin owes us an explanation how the fully indexed ethical value can be general in the minimal sense, required to make intelligible the actions of persons. Acting for reasons, believed by the acting person himself to be constituted by fully indexed values (in the sense of the double relativisation above) is unintelligible.

Dworkin seems aware of the problems with such a position, and I take the distinction he introduces between two types of circumstances of life to be an attempt to remedy these shortcomings. He distinguishes between limitations and parameters of the choice of a good life, and within the latter category, between hard and soft parameters. It is time to briefly consider whether this move helps to deflect the objections from the unintelligibility of fully indexed values, as well as from the impossibility of mistake.

4. Limitations and Parameters of Good Life.

Dworkin's distinction between limitations and parameters of choice of a good life is meant to solve a two-tier problem.

On the one hand, is the problem, that if all circumstances of the choice were considered *limitations* on realising maximum value in one's life, this would imply a transcendent view of impersonal value: its realisation is hindered by the circumstances of a given life. If some of the circumstances (Dworkin dubs them parameters, as distinct from limitations) are taken as *defining* the challenge he is facing, specifying the criteria for a

successful performance in meeting this challenge and thus determining what counts as a good life (or value) for that individual, one accepts an indexed account of value.

On the other hand, the problems discussed above emerge when the challenge is fully indexed to one's circumstances-as-parameters. The remedy is to postulate that some parameters of choice should be taken as *normative*, setting some *external standards* (deviating from the circumstances at the moment of the choice of challenge and specifying what the right circumstances should be) for goodness of a life. The problem with this suggestion is that the normative parameters on the challenge model should still be determined from "*within*" the lives of the persons (they are assigned such a normative status *by the persons themselves*), and thus cannot be external in a sufficiently strong sense.

4.1. Normativity through Parameters?

Dworkin further distinguishes hard and soft parameters. "Hard parameters state essential for a specified performance conditions: if they are violated, the performance is a total failure, no matter how successful it is in other respects." Soft parameters also define an assignment, but they are not each separately essential for the success of performance: failure to meet them can be compensated for by meeting to a higher degree the requirements of another soft parameter or hard parameter[181] of the choice of a challenge. The hard parameters seem to hold the promise of bringing in normativity, as a constraint on the choice of life. They specify the essential conditions for a successful performance/meeting of a challenge: if one fails to satisfy those conditions, one has failed in meeting the specified challenge. However, since one himself determines which, if any,[182] parameters are hard for him, and for how long, as well as judges his own success in meeting them, again the prospect of an external constraint, indeed of normativity disappears, and the possibility of mistake with it. If the determination of the right for the individual person "mix" of realised value - how much one has met the requirements of the many collectively necessary conditions or "parameters" for a successful life, - is thoroughly *indexed* (internal) to him, it could hardly count as a standard, against which mistakes can be judged.

[181] Dworkin (1989a: 70)

[182] According to Dworkin (1989a: 70) most people believe that all parameters that define the success of their life are soft.

Let me also point out that besides the theoretical problems it presents, the picture of thoroughly indexed ethical value rests on an implausible phenomenology. People tend to accept that at least some parameters of their situation are hard, external to them and inflexible, not open to constant redefinition. They define the main features of their identity. Their identity might be multiple, but this would only show that one accepts many parameters as hard, not that one does not accept any.

"Normativity" brought by soft parameters in unintelligible. One could hardly plausibly claim that the aspects of one's situation one oneself decides to define the success of one's life, *confer* goodness on it. Such goodness would be entirely indexed to the given situation, the given person, and his given convictions about what should be the parameters that define it. What the goodness here is, would be unintelligible from any other perspective than that of the person choosing the parameters, in the exact moment and the exact circumstances, with the exact convictions, he is choosing them. Such normativity would be unintelligible for the person himself as well. One cannot form an idea of what would be good, or of value for him, *unless* he were able to see it as a good *independently* of him deciding it is good for him, i.e. as a non-fully indexed, in some respects general good or value. This shows the belief that all parameters defining the success of one's life are soft, to be *incoherent* as well.

It is thus necessary to bring into the challenge model some elements of 'transcendent' value.[183] But since the challenge model is premised on indexed both to one's circumstances and to personal choice of right challenge value, bringing in such elements of transcendent value may render it incoherent. The problem we started with was that unless Dworkin can successfully introduce some external constraint on the otherwise unconstrained choice of value, he could not maintain the crucial for his account of well-

[183] The parts from Dworkin (1989a), where Dworkin discusses the continuity/discontinuity strategies of reconciling the personal with the political (impersonal) perspective, where he criticised the discontinuity strategy, defended by contractarians like Rawls and Scanlon, precisely on the ground that it does *not allow* for justice to have normative force, have been omitted from the chapter in Dworkin (2000). Note especially n.23 in Dworkin (1989a: 34 –35) ' … the interpretive version of the argument [for a political conception of justice] ends back where we began: needing an independent, non-interpretive argument for the categorical force of a political conception of justice.' I take the main problem of his *Tanner* lectures to be precisely how to account for the *normative force* of a conception of justice, *without* either succumbing into a *"transcendent"* account of value, nor falling back within the *discontinuity strategy*. The attempt to tame and surpass the instability of this position is what makes the lectures so challenging.

being distinction between volitional and critical interests. The "softish" parameters did not help here.

4.2. Justice: The Universal Normative Parameter?

Dworkin does have an argument I have not yet considered, which might vindicate the challenge model - the argument from justice as a normative parameter. Justice, according to Dworkin, is a normative parameter, which brings external constraints on the personal choice of a right challenge, without introducing any transcendent value.[184] Justice is at least a soft parameter of our choice: we cannot normally lead good life in circumstances, which are short of being just.

My claim is that though justice may constrain the personal choice of a challenge without introducing some transcendent value, it cannot sufficiently determine the right challenge one is to set for oneself. One criterion of success in this respect is that it should allow us to systematically distinguish between volitional and critical interests. It is again not clear how the introduction of justice as such a constraint helps in this respect.

Firstly, the way justice (just distribution) is determined, is insensitive to that distinction. Let me explain why. The just distribution of resources, (for Dworkin it is an equal one), is determined by measuring the opportunity costs of one's choices for the rest of the community. These costs to others, establishing how much of the community's resources are legitimately one's own, is determined by a market mechanism, where the preferences of all the members of the community set the price of different resources. The market cannot discriminate between preferences, promoting the critical interests from promoting the volitional interests of the members of the community. This is precisely what makes the market the best devise for determining the opportunity costs: here one *should not* discriminate between simple preferences (volitional interests) and convictions about the good life (critical interests). Any attempt to distinguish them from *a different than a first-person perspective*, would involve *paternalistic interference* with the choice of valuable

[184] Dworkin's claim that justice is a normative parameter for all without being dependent on transcendent value, is challenged by William Galston (2001: 611). He claims that Dworkin offers a single conception of justice as constraining the good life in all times and societies, which rules out many substantive conceptions of virtue and the good. He concludes: "If the content of justice is transcendent and if justice is one of the parameters of good lives, then the content of good lives is to that extent transcendent as well. If so, Dworkin is offering a more classical foundational account of ethics and politics than he has yet acknowledged." Galston (2001: 611)

challenge. Thus the market mechanism of determining the just distribution of the communal resources, is set to work *from the perspective of the persons, who are choosing challenge for themselves*, and not from some external point. The result from this process: just distribution of resources, is to serve as a constraint on the acceptable ways of life, bringing in an external (determined by the community as a whole, oneself included) normative element in the personal choice of an appropriate challenge.

It is not clear, however, how the result of a process, in principle *insensitive to the distinction between critical and volitional interests*, can help maintaining it. It is not clear how justice as a constraint can somehow "create" this distinction out of nothing, if the way justice itself is determined is insensitive to it. Certainly, justice does not itself exhaust what people believe is in their critical interest, though surely it sets some constraint on how much one can "spend" on what one believes are one's critical interests. This constraint, however, is purely formal, in effect stating: do not exceed the limits of your fare share of resources. As such it cannot solve the substantive issue of determining the criterion for the correct application of the distinction between volitional and critical interests. It does not show, and is not intended to show, how the good (the critical interests) as distinct from the simply volitional interests, can be derived from the right (the just distribution of resources). It only shows that the right limits the good.

The suggestion that the distinction critical/volitional interests is determined by whatever happens to be the just distribution of resources,[185] could be thought supported by Dworkin's thesis that "justice limits ethics," when interpreted as "the good life for a person is partly *determined* by what are the resources, which are legitimately that person's own." But even if so, it seems not enough for determining the criterion for the correct application of such a central distinction for an account of well-being, as the distinction between critical and volitional interests. Were it to be sufficient, it would turn out that one is not allowed to satisfy those of one's volitional interests, which are not in one's critical interest to be satisfied.[186] If the just distribution of resources sets the

[185] Dworkin's position comes close to this suggestion:

> "Ethical liberals believe that the *character* of people's critical interests depends upon justice: they cannot know, in adequate detail, *what* their critical interests are, until they know, at least roughly, what distribution of resources among them is just" Dworkin (2000: 278, emphases added)

[186] Recall that for Dworkin it is in one's critical interest to have some of one's strong volitional interests satisfied.

dividing line between volitional and critical interests, the satisfaction of volitional interests as such (not as part of one's critical interests) is excluded a priori: one can legitimately pursue only those ways of life, which draw only on one's fair share of community's resources. Since what ultimately determines one's success in life is one's success in one's critical interests, and one's fair share is exhausted for satisfying one's critical interests (by definition, the critical interests are those and only those, satisfied using one's fair share of the communal resources) one is never permitted to satisfy one's volitional interests simply as volitional interests (as *not* being in one's critical interest to be satisfied). This is unsatisfactory, because it is clearly against Dworkin's intentions,[187] to defend both volitional and critical interests taken separately as legitimate concerns both for the individual and for the community. This is a strong reason to insist that the distinction between the two interests be determined by introducing some other *substantive normative* criteria, constraining personal choice of appropriate challenge.

Justice, let us admit, seemed a suitable candidate for playing that role, since on Dworkin's theory it is the only parameter normative for all individuals (even if it is not a hard parameter for all). Were this move to be successful, it would have allowed us to at least initially deflect the argument from the impossibility of mistake One would be mistaken in believing that something is in his critical interests, if it required more than his fair share of the communal resources. It would have also answered the argument from the unintelligibility of a fully indexed value - justice would be constraining the choice of challenge, thus introducing an element of a not fully indexed value.

However, this argument would have only initially met the objection from the impossibility of mistake - normative constraint on personal choice of value, such as one's just share of resources, would not be sufficient to distinguish between valuable and valueless use of one's share. One's life could conceivably still lack value even though one did make use *only* of one's fair share of resources. Thus, even if the argument from the justice as a normative parameter is sound, one would still need a further normative constraint on personal choice of value, in order to meet the objection from the

[187] Dworkin (2000: 245) is careful enough to point out, that though the defence of his type of liberalism depends on concentrating on the critical as distinct from volitional interests, the liberal community should be concerned with improving its members' life in the volitional sense as well. This is so, because people have reason to care for their life in both the volitional and the critical sense.

impossibility of mistake. I believe I have already discussed all the candidates for this role proposed by Dworkin, and found no convincing argument for meeting this objection.

Given the arguments for the insufficiency of justice as constraining the personal choice of challenge non-indexed value, and given the absence of a more plausible candidate for that role, one has to admit that Dworkin has failed to establish the coherence of the challenge model of critical well-being. To repeat, the problem is, that on the underlying that model indexed account of value, value is thoroughly indexed to a personal choice of value, which, when left unconstrained, makes choice a sufficient condition for value. Accordingly, the distinction between volitional and critical interests is blurred: for both endorsement [choice] is a sufficient condition.

5. Cultural Paternalism and the Endorsement Constraint

One of the reasons I opted for the strong interpretation of indexed ethical value was that only on it Dworkin's challenge model of well-being rules out *cultural (substitute) paternalism*. This paternalism aims at removing worthless options and their substitution with worthwhile ones, so that people are aided in their choice of valuable life. Acceptance of such paternalism, according to Dworkin, would show deep misunderstanding of his intuitively plausible model of well-being. But why is cultural paternalism believed to be hostile to good life? Is there something intuitively objectionable in having to choose one's life from a list of paternalistically pre-selected options - cultural paternalism's aim?

The intuitive reaction against cultural paternalism has to do with the thought, that one's success in life depends crucially on one being allowed to take the credit for defining the challenge of one's life according to one's own convictions. If one's choice is limited and defined by a pre-selection of the options to be available, one is denied the opportunity to take the credit for one's choice of way of life. Crucially here, the credit is greater, where the risk of mistake is present. The problem with such paternalism then is that one cannot be given the full credit, and accordingly, be held responsible, for the choice of one's life.

This is not immediately obvious. Consider that on this most sophisticated form of paternalism, besides the requirement that one choose a *worthwhile* (component of) life, it is also important, that one chooses one's way of life out of one's own convictions that it

is worthwhile. One only is denied the opportunity to form wrong convictions, given that, ideally, one is not offered any worthless options at all (one chooses from a deliberately tailored list of objectively valuable options). Thus the ethical integrity of persons is preserved: they are never made to act against their own convictions as to what is good for them. There should be something else objectionable in cultural paternalism, not directly connected with the threat to the priority of the ethical integrity of the persons. May be it could be found in those aspects of endorsement that go beyond the requirement of preserving the ethical integrity of persons.

In the analysis of the definition of endorsement, I claimed that Dworkin wants to go beyond the understanding of endorsement as a positive cognitive/affective response to the valuable components of one's life. It should now be clear why. If one accepts such a definition of endorsement, (1) cultural paternalism would not be excluded, and (2) endorsement itself would not be taken as *constitutive* of well-being. For cultural paternalism, as we saw, does not deprive one of the possibility of forming such positive evaluative/affective attitude towards the pre-selected options, and it does not threaten the integrity of persons. They could not normally (under conditions of successful cultural paternalism) form convictions, which could hinder their genuine satisfaction with the available worthwhile options.

The important difference between cultural paternalism, and the challenge model, is that only on the latter one could have failed to endorse the worthwhile life, and in that sense the credit for endorsing it is one's own. The success of one's life is truly of one's own making, only when one can take the *full* credit for choosing a worthwhile life, and living up to its standards. Notice that one can still take a credit in leading a good life, even if he did not choose it entirely on one's own. Thus, the success of one's life is a matter of degree, which is partly determined by the degree to which one is responsible for its choice (given that the choice is right, of course!). One cannot have a successful life, however, if one cannot take even in principle any credit for choosing it, since all the options were (equally) worthwhile, and one had only to stretch his hand and grab one of them, irrespective of which one exactly he actually ended up choosing.

Thus the objection against cultural paternalism is that on it *endorsement* of a worthwhile component of one's life loses its significance. Endorsement becomes an automatic

response. If one cannot in principle in these circumstances be mistaken, it does not make sense to deliberate and form convictions – they can never be mistaken! Endorsement under such conditions is a response to an independent value, which could only fail to emerge, were the individual to suffer from some cognitive or psychological defects. In short, endorsement makes sense only in situations, where it is possible for one to be mistaken about which life is worthwhile. Only there one could truly be *the author of one's life, and the ultimate source of its value.* The full credit for the creation of value cannot be entirely one's own in circumstances, when in principle no mistakes could be made: one is deprived of the challenge involved in creating this value.

Precisely at this point I find Dworkin's position puzzling, even paradoxical. We already established in a previous section that this strong view of endorsement (not simply a positive response to the components of one's life, nor even identical with priority of integrity), renders implausible the challenge model of well-being precisely on the ground that mistakes and criticism are not possible there. What Dworkin finds objectionable in cultural paternalism from the point of view of the challenge model, however, can most plausibly be described exactly as "impossibility of mistake." This is hardly coherent.

One could object that I have misunderstood, or misinterpreted the reason Dworkin finds this type of paternalism hostile to his account. His objection might be not that the challenge of choosing one's life is blunted when wrong options are removed from the choice set. Rather, he might instead object that the challenge is made less "challenging" because of the reduced diversity of the options: a possible effect of cultural paternalism.

But this could not be a principled objection against cultural paternalism - the aim here is not to reduce diversity, diminishing the opportunities for an interesting challenge. If it has that effect, it at best is an undesirable side-effect, unintended consequence. Were there to be such undesirable consequences, this could justify the prohibition of paternalism out of efficiency considerations, not out of principle. The aim of cultural paternalism, recall, is to remove the bad options by substituting them with worthwhile ones. Precisely this is found objectionable from a Dworkinian perspective. And the best explanation why is such substitution objectionable is that endorsement makes sense only when the possibility of mistake is present - there is no challenge, when the possibility of mistakes is removed.

(It is a further question whether endorsement, as understood by Dworkin, allows for the possibility of mistake - as already pointed out, his position does not seem coherent here.)

If this is a sufficient ground for rejecting cultural paternalism, will it also be a sufficient ground for demanding from political authorities to supply bad options - in the extreme and unlikely situation, when they are altogether missing?[188] In this way, the opportunity for a choice of a more interesting challenge would only be enhanced. If this may sound absurd,[189] why is not absurd Dworkin's objection against cultural paternalism?[190]

Let me admit, for the sake of the argument, that my interpretation above of the intuitions behind the rejection of cultural paternalism is not faithful to Dworkin's own intentions. After all, he has allowed for short-term paternalistic measures, as well as for liberal education, which can be taken as a form of cultural paternalism. His main objection is against longer-term paternalism: it may weaken the persons' *capacity for critical reflection*, thus rendering the subsequent to such paternalism endorsement *non-genuine* (not based on convictions that are truly one's own). This is the charge that cultural paternalism is manipulative, piping the thoughts of someone else into the thoughts of the persons. The convictions, formed in this manipulative way, even if not conflicting with the ways of life, available in those circumstances and adopted by the persons with those manipulated convictions, would not enhance the integrity of the persons, since they would not be their own convictions in the first place.

If this is Dworkin's objection, I find it even weaker. It is not clear in what ways exactly would the availability of more good options, and the reduction of bad ones (1) threaten one with forcing him to form non-genuine convictions, (2) threaten one's capacity for critical reflection.

---

[188] Raz's response to the objection against his paternalism that it removes evil options thereby preventing people from freely avoiding them is that any governmental action is unlikely to remove all the evil or worthless options. And even if some of them were, the capacities needed to avoid them, which could not be developed if they are indeed missing, would then be worthless. Raz (1986: 380-381).

[189] Some, following Mill, may not find this suggestion absurd. They might think that mistakes are good for developing one's critical capacities for reflection, may be conducive to discovering the truth, etc. So the government may have the duty to supply bad options when they are altogether missing. This is beside the point here: if it is indeed the case that all bad options are removed, one would not need such capacities either. Not that this is a likely scenario. But then we have Raz's response: there is nothing bad in eliminating bad options– enough of them (fortunately for the critical reflection fans and unfortunately for the die-hard perfectionists) will remain anyway.

[190] My argument, clearly, is a reductio ad absurdum. As such, it reveals and challenges certain basic features of the criticised view. Obviously, for the purposes of refuting the latter, Raz's own argument – bad options are unlikely to disappear altogether, is good enough.

Concerning (1). Consider the situation (a) where one is provided with more valuable options (substituting the available up to this point, non-valuable options), against the background of *one already having formed one's own convictions*. In this case, one would be required, even forced, to leave behind one's wrong convictions. This would presumably be unacceptable, because one's new convictions would most probably not be genuine, or at least would be suspect: they might have been produced by an adaptive mechanism of adjusting to the manipulated environment. Under such conditions of coercive paternalism, one's ethical integrity would be compromised.

The situation (b) is when we already have cultural paternalism. It is different from situation (a): one is not "forced" here to abandon one's wrong convictions. Rather, one is "manipulated" in such a way, that one is not capable of forming them in the first place.

If one is to criticise such a practice (in (b) case), on the ground that it is manipulative, however, one needs to be able to give determinate value to counterfactuals of the sort: what would have been A's convictions, were the circumstances of A's conviction-formation to be free of any manipulation, i.e. be determined "ideally." This is an impossible exercise. Firstly, any situation would turn out according to this test, to be "manipulated" in a minimal sense: certain options are available, others absent. There is no absolute level ground, from which one is absolutely free to form one's own convictions in a "vacuum," with all the influences of available/absent options successfully screened out. Secondly, even if this problem could somehow be solved, it will still be indeterminate what one's convictions would be, were one to abstract from *all* of one's present convictions. There would be nothing, which could give a determinate value to such counterfactual convictions, if they are entirely cut from any person's identity-defining actual convictions: one could freely attribute any convictions to such an entirely disentangled person. In sum, the manipulation objection against cultural paternalism, in the form discussed here, is not conclusive.

Concerning the second (2) threat: if there is something objectionable to the practice of cultural paternalism, it should be connected with the threat it presents to one's capacity for critical reflection. The capacity for critical reflection is essential if one is to be responsible for the choices one makes: it is actively exercised in circumstances, when one can be held responsible for those choices. A person in entirely manipulated circumstances

cannot be held responsible for the choices he makes. Moreover, it presumably hardly makes sense to say that one is responsible for the choice of an option, if all the available options are pre-selected in such a way as to be valuable (ideally, the cultural paternalism's aim).

In short, one seems to need the capacity for critical reflection only in circumstances, when (a) one can make a mistake – can form wrong or unsuitable convictions, and (b) one can be held responsible for that mistake. It should be obvious, however, that once we have reached this point, we run against the objection that it would be absurd to require from the state to supply one with bad options (so that it could make sense that one can be mistaken in his choice, and can be held responsible for a possible mistake), in order to keep one's capacity for critical reflection in good shape.[191] This result would not be absurd, were the capacity for critical reflection to have some *intrinsic value*, independent of its possible contribution to identifying the right choices, the right challenge one is to face. The case for the intrinsic value of that capacity is connected with the case for ECT in its strongest interpretation as requiring that the value for the person of any component of his life is created by his own endorsement of that component.

I have already argued that there is no plausible case for accepting this view of endorsement as being sufficient for well-being. It goes beyond life-satisfaction (endorsement necessary, all things considered)[192] or beyond attributing priority to ethical integrity (endorsement constitutive of well-being) condition. To the extent ECT is defensible, I have maintained, it is exhausted by the weaker interpretation of its being a positive cognitive/affective response to some valuable component of one's life, and requiring the priority of personal integrity.[193] Such endorsement, as we saw, is compatible with non-coercive, cultural paternalism.

---

[191] Recall Raz's response that the complete removal of bad options is anyway an unlikely scenario, so my arguments in the text above are somewhat overdone.

[192] Endorsement in this weak sense is classified "necessary, all things considered" (failing to endorse a component of one's life might bring a drop in one's life satisfaction, not outweighed by the presence of this component in one's life. It is not constitutive for the value of a component. Wilkinson (1996) classifies "endorsement as necessary, all things considered" as strong additive view of endorsement, to be distinguished from the weak additive ("endorsement is necessary, other things being equal") and the constitutive ("endorsement is constitutive of well-being") views.

[193] For a detailed discussion of Dworkin's arguments in support of the priority of the personal integrity, see Wilkinson (1996). This author's claim is that Dworkin fails to rule out paternalism, because he fails to defend the priority of personal integrity. This might be right. But this is not the main reason why Dworkin

This detour into an alternative explanation for the objection against cultural paternalism ended the same way: guided by its logic, we are again committed to the absurd demand that the state supply its people with bad options.

6. Conclusion: Beyond Endorsement

I have failed to identify a convincing rationale, present in Dworkin's discussion, for dismissing cultural paternalism, while maintaining that endorsement is only a necessary, but not sufficient condition for well-being. Cultural paternalism could be ruled out if one accepted the challenge account of well-being, but only if it is premised on the strong indexed view of value. I found the latter implausible, on the ground, that it leaves unexplained the possibility of mistake. On it, value is *created* by personal choice, and accordingly, one cannot be mistaken about value. The further problem with this suggestion was that it proves too much: endorsement becomes a sufficient condition for well-being. A defense of endorsement as a necessary but not sufficient condition for well-being has not been provided by Dworkin.

As a way of concluding the discussion of ECT, let me mention that the unjustifiability of even coercive paternalism need not depend on the plausibility of this thesis even on its weak interpretation. If coercive paternalism is not justifiable, this might be established on other grounds altogether. Raz, for example, conditionally admits the justifiability of coercive paternalism, though he does argue for the importance of endorsement: one's well-being is enhanced by acting on the worthwhile goals one has, and one cannot forcibly be benefited by imposing on one goals, since to benefit him, those goals normally have to be adopted, endorsed by him.[194] The argument Raz advances here does not establish (nor was it meant to) that endorsement is a strictly necessary condition for well-being. One could be benefited, one's well-being enhanced in other ways than through one's goals – success in one's worthwhile goals is only one of the most important contributors to one's well-being. And using the coercive apparatus of the state may be a legitimate way of guaranteeing some of the other conditions for well-being.

---

cannot rule out cultural paternalism. As I show in my text, I do not believe that priority of integrity is compromised by cultural paternalism: even if Dworkin is right about integrity, he might be wrong about this type of paternalism.

[194] Raz (1986: 291-292)

However, the paternalistic use of coercion in addition to "being undertaken for a good reason, sufficient to make reasonable a partial loss of independence," has also to "come from the hands of someone *reasonably trusted* by the coerced."[195] A government that guarantees full citizenship to its subjects can enjoy their reasonable trust and may permissibly coerce them for their own good only - supported by right reason.

This additional, rather innocuous-sounding condition of reasonable trust, is very powerful: it rules out coercive *moral* paternalism, since it is arguable that one cannot reasonably trust a government that paternalistically coerces one against the moral convictions, underpinning one's way of life, thus denying their validity. Such government would be denying him full citizenship – he has reason to doubt that his interests and well-being are indeed taken into account, or given the weight they deserve, in deciding public action, directed against essential aspects of his way of life. Distrust is the reasonable answer here.

What is normatively important, notice, is again the subjective view. It is reasonable to distrust an agency, which in coercing you against your moral convictions, declares them and your whole way of life worthless: it will naturally *seem to you* that in doing so, it does not take your interests and your well-being seriously. The reasonableness of the distrust, however, is independent of whether you are wrong and the coercing agency right. Raz does not embrace subjectivism about value and well-being in order to reject coercive moral paternalism, as Dworkin (as I show in this chapter) does. Raz does not even argue, that endorsement is either sufficient or necessary for well-being: a component of one's life can contribute to one's well-being even if its value was not endorsed by one. Though he agrees with the view that "under all conditions a good and successful life is one of willing and whole-hearted engagement in worthwhile relationships and pursuits, and such whole-hearted commitment to one's life is incompatible with that life being coerced or manipulated by others,"[196] he does not believe that these considerations rule out paternalistic use of coercion.[197] The case he builds for ruling out coercive *moral* paternalism depends on other than endorsement

[195] Raz (1996: 122)
[196] Raz (1996: 121)
[197] He thus identifies these considerations as "one source of the persisting popularity of belief that governmental decrees are legitimate only if self-imposed, i.e. only if endorsed by the general will of all their subjects." Raz (1996: 121)

considerations: it is reasonable to distrust government coercively preventing one from acting on one's moral convictions, since it denies one full citizenship with all it implies.[198]

This conclusion is important for my purposes. One of the hypothesis I test is whether the fact that authority may act against the considered convictions of its subjects should impose limits on its claims to obedience, and on law's claim to comprehensive supremacy over all other normative domains. Dworkin's attempt to justify such limits by defending the endorsement constraint thesis was found wanting. This does not mean that the thesis cannot be defended, even if not in this strongest form. But one need not do that to show that there should be limits on the claims authorities make to impose their own views and determine the lives of their subjects, disregarding their considered convictions. The use of coercion by authority rules out such impositions, especially when the moral, fundamental convictions of their subjects are concerned. In the following chapter I show that there are even stronger limits on the comprehensive claim to normative supremacy political authority (through law) necessarily makes.

[198] "However, if it pursues coercive moral paternalism against me it will, by definition, be preventing me from following my way of life, and it denies, in a purported exercise of its authority, the validity of propositions I hold true and which underpin my way of life. If it does so, however, it denies me full citizenship." Raz (1996: 127)

Chapter Five

The Normative Supremacy Claim and the Autonomy Condition: A Defense of Agent-Relative Reasons for Action

1. Introduction: Autonomy and Agent-Relative Reasons for Action

The recognition of the validity of agent-relative reasons seems necessary for the soundness of what I label a reason-based justification for liberal-democratic authority. The reason-based account of political authority conditions the justification for authority on its capacity to provide its subjects with valid protected reasons for action. This is generally the case when the putative authority brings improved conformity to subjects' own, independent of authority reasons. Thus the validity of the protected reasons authority provides, is not derived from their actual or hypothetical acceptance by the subjects. Rather, it is that they meet the requirements of Raz's Service conception, which constitutes their validity. Whether the requirements of the Service conception are met, and thus the protected reasons - valid, will depend on at least two considerations: the types of reasons subjects have independently of authority and on the capacity authority to be the agency, through which improved conformity to them is brought about.

Thus, an exploration of the reason-based account of liberal-democratic political authority seems to require closer look at the types of first-order reasons for action subjects have and their relation to the purported authority – is it the case that liberal democratic authority is particularly suited to bring improved conformity to the most important reasons its subjects have? Or may be there is some merit in the way[199] the conformity with the first-order reasons is achieved, which bestow validity on the protected reasons the authority provides to its subjects. In this text I explore the first route – I focus on the type of first-order reasons. I admit that the two routes might be necessarily interdependent. It might well be the case that, given the types of first-order reasons the subjects have, they ask for a particular way of decision-making (ways of deciding on the

[199] For example, employing certain procedures for aggregating the opinions of the subjects as to what would help them most, or to the highest degree, to conform to their first-order reasons. Such procedures would be conducive to authority issuing directives, which are actually more likely to provide the subjects with valid protected reasons, than would any alternative way of decision-making.

types of directives authority can issue, if it is to be a legitimate authority).[200] This latter issue will concern the procedural aspects of the legitimacy of an authority, to which I will turn in the final part of my thesis.

In this chapter, instead, I will test the hypothesis that a strong case for justifying liberal-democratic political order can be made, if it is true that subjects have agent-relative first-order reasons for action along with their agent-neutral ones. The connection with Raz's autonomy condition is obvious. Recall that the autonomy condition is a necessary constraint on the test of legitimacy provided by Raz's NJT: only if it is the case that it is more important to decide certain issue correctly rather than act on one's own, is following authority's directives justified. However, if certain types of reasons either (1) do not allow that improved conformity to them is achieved through acting for (complying with) some other reasons, or (2) do not allow for the maximising logic inherent in the notion of improved conformity to reasons, or (3) make it the case that improved conformity to them does not matter, and if these types of reasons sufficiently populate the space of subjects' reasons, the scope for the operation of NJT as a test of legitimacy will be severely limited. This will threaten its position as the central, main test for legitimacy. This conclusion, however, will be contingent on the empirical fact as to how wide-spread these (yet only hypothesized) types of reasons are. Notice, however, that the conclusion to be possibly drawn from the success of my arguments here is much stronger than this rather limited, contingent claim.

My contention, motivating this chapter, rather, is that if there are such types of reasons (agent-relative ones of different sorts), limiting the scope of the NJT, the claim authority necessarily makes to normative supremacy over all other normative domains, will necessarily be false. It is so, because in the case the nature of those reasons limits the justified exercise of authority, it is those reasons, and not authority that determine this limitation. What, then, may characterise a liberal-democratic type of authority is that it respects such reasons, and that it refrains from making the utterly implausible claim to normative supremacy over all other normative domains.

[200] The presumption here would be that this and only this way of decision-making would satisfy the requirements for justifying authority, mentioned above.

This is how I come to my main topic in this part of my thesis: *agent-relativity and its importance for a reason-based account of the authority of a liberal-democratic political order*. My aim here is first to offer a working definition of the concept of an agent-relative reason for action. In addition, I attempt to advance some arguments in defense both of the usefulness of the concept and of the possibility of having valid agent-relative reasons for action.[201] I also offer a brief concluding discussion about the autonomy condition, (and the two arguments for it I explored in this part of my thesis – the endorsement constraint thesis and the presence of agent-relative reasons for action) and its relation to one of political authority's (or its law's) essential features, according to Raz – its claim to normative supremacy. My conclusion is that this claim being an essential feature of political authority and its law – not only of liberal-democratic, but of any type of political authority, cannot be reconciled with the central tenets of the NJT and the Service conception of legitimacy more generally. The reason is that the Service conception necessarily conditions the legitimacy of law and political authority on ultimately being morally justified (or required by practical reasons). Since it is morality, or practical reason, that ultimately justify authority, authority's claim for supremacy over all other normative domains is necessarily false, and thus obviously implausible. Making obviously implausible claims could hardly be an essential feature of authority. If it is in our concept of political authority, that it necessarily makes such a claim, then it is the account of justification, on which this claim turns out to be necessarily false, that is inadequate. If it is the instrumentalist account of legitimacy that seems central to our concept instead, then the claims authority is thought to necessarily make should be modified. In either case, something should be given up: either the claim itself, or the instrumentalist account of legitimacy should go. The two do not seem to go well together.

[201] After their introduction by Nagel (1970: 90) (as subjective and objective reasons and values) and Nagel (1980: 77-139) (already as agent-relative versus agent-neutral reasons for actions), the terms were discussed by Parfit (1984: 104) and further clarified in Nagel (1986: 152-153), which sparked an extensive debate over that coherence and the use of the concept of agent-relativity. For useful discussions, see Scheffler (1982), Kagan (1991), and the essays in Scheffler (1988), among others.

I start the discussion with defining agent-relativity.[202] Next, I try to discern what is the motivation behind the recognition of agent-relative reasons – and it seems to be the divergence between the neutral value of an option, and the value of the same option to the agent. Thirdly, I distinguish two possible explanations of this divergence: agent-relative, and agent-neutral, and try to show that the latter is unsatisfactory. My conclusion at this stage of the argument is that the concept of agent-relative reason for action plays an important role in explaining the often-observed divergence of value. Though this argument, I believe, provides strong support both for the plausibility and usefulness of this concept, as well as accounts for how such reasons could be valid, there are additional problems with this concept that need exploring. Thus the fourth step is to see, whether the agent-relativist position is not weakened or jeopardized because of its problems in meeting the requirements any account of reasons for action needs to meet. I admit that this is a serious challenge for the agent-relativist, but I claim it could still be met. The last problem I discuss is whether the defense of agent-relative reasons for action does press one to abandon the plausible reason-based account of autonomy, in favour of a will-based one. I try to show why this need not be the case. At the end of this part of my text, I suggest that if the validity of agent-relative reasons is recognised, this would ask for stronger (than Raz suggests) constraints imposed by the autonomy condition over the liberal-democratic political principles, enabling authority better to serve the interests of its subjects. I also elaborate in more detail on the far-reaching conclusion, indicated at the end of the last paragraph.

2. Agent-Relative Reason for Action Defined: Structure and Main Types

What is the meaning of the statement 'A has an agent-relative reasons (ARR) to F', and how does it differ from 'A has an agent-neutral reason (ANR) to F'?

[202] Given my interest in exploring what first-order reasons for action an agent can have, I will be concerned to define agent-relativity as concerned primarily with agent-relative *reasons*, though it is clear that these reasons cannot be taken apart from their interconnectedness with agent-relative values, on the one hand, and with agent-relative theories, on the other. This is suggested by Jonathan Dancy in "Agent – relativity - the very idea," in Dancy (1993: 208).

A plausible explanation for the difference between ARR and ANR for the same act F, is that there is a divergence between the value of the act impersonally judged (constituting the neutral reason ANR to F in the latter statement), and the value of the act for A (constituting the relative reason ARR in the former statement). More formally:

V*A (FA) # V (FA) – where V*A is the relative value for A of A F-ing, and V is the neutral value of A F-ing.[203]

One and the same action (taking care of one's own children) can have *both a neutral and a relative value.*[204]

In the former case the neutral value of the option (the prospective action) can provide *two types of neutral reasons* for action – action-reasons (the value is in the performance of the action itself) and outcome-reasons (the value is in the consequences of performing that action).[205] The neutral action-reasons apply directly to those who are in a position to perform the action - take care of their own children (they would not be able, if they were not parents). The neutral outcome-reasons, on the other hand, can apply to any agent, who is in a position to help the parents take care of their own children.

I suggest that when the *value* of the action is *relative to the agent A* (this is the case of V*A(FA)), it *constitutes an agent-relative reason for A to F*: I have a reason to take care of my own children, which is different (I should probably add – *different in kind*, but this is to be first established) from the neutral action-reason I as a parent have to take care of

[203] Here I again use a formulation given by Dancy (1993: 200).

[204] This position can be accepted both by an agent-neutralist and by an agent-relativist. For the suggestion that a neutralist can accept it, see Raz (1999: 64) "While the fact that competence on the piano is John's goal does not affect the value of such competence, it does affect its value to John." The divergence of value is explained by the appeal to the presence of goals: once the agent "has made something his goal, it acquires special importance for him. He has a *reason* to pursue it that he *did not have before.*"(Raz 1999: 64, emphasis added)

[205] I already discussed this distinction in chapter 2 of my thesis. Raz follows Parfit (1984: 104) in introducing the doing-happening distinction. He attempts to use it for accommodating certain agent-relative concerns within an agent-neutral moral framework. See Raz (1986:279-281). For an illuminating discussion of the difficulties attending the attempts to specify the logical form of the agent-neutral/agent-relative distinction, see McNaughton and Rawling (1991). These authors believe that the only way to save the distinction from the "consequentialist vacuum cleaner" (if it is *valuable* that A act on his agent-relative reason to F, then B should ensure that A act on it, and this move swallows the distinctness of the agent-relativity in all but a small group of cases, where B cannot ensure that A act on his reason unless B is identical with A) is to distinguish sharply between the deontic and the evaluative; see McNaughton and Rawling (1991: 180). Accordingly, my own way of delineating the distinction in terms of the divergence of value would be unsatisfactory, unless I manage to show that the value of F for A instead of being enhanced is altogether lost when B tries to ensure that A acts on it. Here we could find the obvious link with the ECT we discussed in the preceding chapter.

my own children. Even if the neutral value of F would be promoted (more people would F) if I were not allowed to F (because I happen to prevent many of them from F-ing), I still have a relative reason to F, deriving from the value for me of me F-ing, which is different from and not reducible to the neutral value of me F-ing.[206]

This divergence of value is a divergence of objective value – it is not simply a possibly justified subjective distortion of the valuation of F, due to the unfortunate partiality of the personal perspective.

Before going into discussing whether this observed divergence of value has implications for the validity of agent-relative reasons for action, let me introduce the distinction between different types of agent-relative reasons. Next, I will need to see whether the above formal presentation of the structure of first-order reasons for action is capable of accounting for all of these types

An example of an agent-relative reason of *autonomy* would be:

> 'The value for me of me writing a text on agent-relativity is different (given my *goal* of writing a thesis on a connected topic) than the impersonal value of me writing this same text: the relative value of this option (writing a text on agent-relativity) provides me with an agent-relative reason of autonomy to write it.'

It might be impersonally indifferent, whether I write it or not. It might even be much better impersonally judged, to have somebody else write it, using the same external resources and performing better, but the value of action for me gives me a reason to perform it – and this reason seems to be an agent-relative reason of autonomy.

---

[206] A question, not fully addressed in this text is: whether the agent-neutral and agent-relative value distinction can be maintained within a broadly teleological framework. The difficulty is: if a teleological interpretation of the relative value is accepted, this might allow that the dis-value of me breaking my promise is outweighed by the greater dis-value of me breaking many more of my promises in the future, unless I break this promise now. This is counterintuitive: me breaking a promise is wrong, irrespectively of any badness of its occurrence. So, the distinction deontological agent-relative/reasons of autonomy can be spelled out thus. The divergence of value in the case of deontological reasons is a divergence between teleological (where the value or the good is whatever is to be maximised) and non-teleological value (where value, or the good, should not be maximised). The divergence in the case of reasons of autonomy, on the other hand, is divergence between two types of teleological value (one from an impartial point, the other- from the personal perspective of the agent, where both are to be maximised). This difference between deontological/autonomy reasons may threaten my analysis of agent-relativity in terms of the divergence of value, unless a further distinction within the category of value is made – to be promoted or to be respected only; Scanlon (1999). Arguably, a similar function plays the distinction in Raz (2001b) between respecting value and engaging with it.

The agent-relative reasons of *partiality*, stemming from personal commitments and relations, exhibit the same structure.[207]

The merit of the formal presentation above I take to be that it may capture the idea behind *deontological* agent-relative reasons as well, (allowing it to cover the most commonly discussed types of agent-relative reasons, and thus to exhibit the common structure of agent-relativity):

'dis-V*A (FA) # dis-V (FA)' – the dis-value *for me* of me breaking my promise is different in kind and irreducible to the impersonal dis-value of me breaking my promise. The dis-value for me of me breaking my promise provides me with a deontological agent-relative reason for not breaking my promise.

Here the dis-value of the act for me (which provides me with a relative reason against performing it) is not due to its being disallowed by a project, or a goal, I happen to have. Irrespectively of whether I endorse the prohibition against lying, killing, etc., I am required to refrain from such acts, even if their performance would minimise the occurrence of such acts at the current time-slice by other agents, or in the future both by other agents and by myself.

3. Divergence of Value: Agent-Relative or Agent-Neutral Explanation?

The divergence between the neutral value of the option and the value of the option to the agent I take to be the central case for establishing the validity of agent-relative reasons.

3.1. Divergence of Value?

The thesis about the divergence of value has an important place in the contemporary debates about well-being and its relation to objective value. The well-being theorists defend different theories of well-being, but in all of them the distinction well-being/objective value makes sense: the question asked is usually which are the values that contribute to one's well-being – to the 'value of one's life for' the person concerned. Objective list, preference-satisfaction, and hedonist accounts of well-being give diverging

[207] See the discussion above of the example of parents having both neutral and relative reasons to take care of their own children.

answers to this question, but for all of them it does make sense: they do recognise that neutral 'value' and relative 'value for' may in principle diverge.

Some theorists[208] deny that this divergence has normative significance: they deny that the divergence involves more than an empirical relativisation of the neutral value: 'for' in 'value for' means nothing more than 'value, occurring in the life of.' Thus whatever normativity is involved in 'value for,' it is entirely accounted for by the universal normativity of 'value.' The argument offered is more than simple – when we ask ourselves what is of value in our lives (value for), we try to transcend our immediate desires and ask ourselves whether our most fundamental desires are worthy ones: whether they have 'value.' If there is no place here for 'value for' as normatively distinct from 'value', there is no place for agent-relative reasons as distinct from agent-neutral ones either. This simple argument, however, does not establish much.

There is a compelling argument in favour of the divergence of 'value' and 'value for' and the normative significance of the latter. If the concept of the good *for* an agent is abandoned, then self-sacrifice: sacrificing one's own good for the good of another or the good overall, becomes impossible.[209]

The quick reply that self-sacrifice as so defined need not be possible – acts of self-sacrifice can plausibly be re-described as acts involving ability to resist strong natural impulses in the service of the good, should be dismissed since it denies individual well-being normative significance. This certainly is going too far: from the argument that when we ask ourselves what is of value in our lives, we ask about 'value' and not about 'value for,' (recall the simple argument of the neutralist rehearsed above), it simply does not follow that individual well-being has no independent normative significance. The

[208] Regan (2004: 202-230) is an example of such theorist, who follows Moore (1903) in denying that the notion 'good for' is normatively fundamental, competing on an equal footing with 'good simpliciter.' The references in my text are to Regan's article.

[209] Regan (2004: fn 52 at 224). Regan attributes the self-sacrifice argument to Overvold (1980), made more popular by Darwall (2002). The response Regan gives (see in the body of my text) to this argument was challenged already in Raz (1989) (see the footnote that follows) and it is strange Regan does not use the occasion to answer that counterargument in an article on the same issue, included in a volume devoted to Raz's work on this topic. It is a common place to use the self-sacrifice argument in discussions of agent-relativity as well (Dancy 1993).

normative significance of well-being does not, nor can it depend on the irrelevant fact whether this (well-being) is what individuals desire.[210]

There are theorists, who accept the possibility of divergence of value, but deny that it implies agent-relative reasons.[211] I argue against such a position, by trying to establish that it is inconsistent.

Lastly, there are those, who start from what I want to establish – they accept that there is a difference in the reasons our projects, goals, relations, etc. give us, compared to the reasons we have to value things that are not so closely connected to us. They are not, however, willing to attribute this difference of reasons to a difference in their value.[212]

Part of the resistance to the thesis about the divergence of value, is due to the apparent instability of the claim that a person who has a commitment, goal, or a relationship is somehow liable to see their value as higher than those of other peoples' commitments, goals and relationships. By universalizing this claim, the argument goes, we end up with the absurdity of the value of *all* projects, commitments, goals, relationships, etc. being *above the average*. Such theorists try to avoid this obviously absurd conclusion by drawing a distinction between neutral value and importance to oneself of one's goals. One need not regard the value of one's project as higher than that of somebody else in order to admit that that project may matter more to him just because it is one's own project.

I do not believe that the distinction between *value* and *value for* has the above-described implications: value and value for need not be compared, at least within certain limits:[213]

---

[210] There is a sophisticated discussion of this and related points in Raz (1989: 1212-1217), where he responds to Regan's argument that if people do not aim at their own well-being (a view they share), this shows that they do not value their well-being, and that, further and crucially for Regan's purposes, their well-being is not a distinct value. My argument in the text was influenced by Raz's discussion.

[211] Raz, for example, distinguishes between personal and impersonal value Raz (2001b: 83-84), roughly corresponding to the distinction value/value for in my text, but he does not believe this implies something like agent-relative reasons for action, Raz (1986: 277-284). He claims that agent-neutral action-reasons, provided by personal commitments, relationships, projects, etc., exhaust what is believed characteristic of agent-relative reasons. Nevertheless, he maintains that our decisions, commitments, relationships, etc. have normative significance – they create new reasons for action of a special kind. (Raz 1978a) This Razian position – that reasons not only can be discovered – in the independent value of an option, but can also be created by an individual choice, is criticised by Scanlon (2004: 231-246).

[212] This I take to be the position of Samuel Scheffler (2004: 250 –251).

[213] This, of course, is Raz's idea that decisions and commitments have normative significance in that they provide new reasons for action of a special kind – protected reasons for action, with an exclusionary

so the unstable position with all commitments, goals, projects, etc. having above the average value, need not arise. Moreover, I do not see too much difference in saying: 'I have a reason to F because F-ing and not P-ing matters to me, even though P-ing has an equal to F-ing value', rather than saying: ' I have a reason to F because F-ing has a special value for me, even though P-ing has an equal value. If comparison to P-ing were necessary to explain why F-ing is a legitimate course of action, it will certainly be involved in the former, if not in both (because of the exclusionary element in the latter, at least on Raz's account).[214] There is a further advantage in the 'value for' formula over the 'importance for' one. The dependence of reason on value is preserved: the reason for A to F resides in F-ing's value for A, while it is not immediately obvious how the fact that F-ing is important to A gives A reason to F, if F-ing and P-ing have equal value (to A and to all other agents). Both considerations suggest that the divergence of value thesis is an acceptable starting position.

The divergence of value is usually taken to be due to the fact that performing F is or is not part of A's projects, goals, commitments, relationships, etc.[215] If this were the only explanation for this divergence, it is unlikely that it can explain the validity of all types of relative reasons. It can explain the relative reasons of autonomy and partiality. A difficult question for the agent-relativist theorist (and for the neutralist, who is to be able to account in neutral terms for the plausible intuitions behind all types of agent-relative reasons) is to explain what can make the divergence of value in the case of the deontological agent-relative reasons intelligible.[216] It cannot be due to the presence or absence of goals, projects, commitments etc., because what precisely distinguishes the deontological constraints from the other relative reasons is that they are unconditional. They do not depend on such more or less contingent and optional[217] factors as the presence of goals and projects.

---

element in them. The exclusion means that, within limits, there should be no comparison between the new reason provided by the commitment, decision, etc., and the pre-existing reasons.

[214] See the immediately preceding footnote.

[215] This is the position of both Dancy (1993: 200), and Raz (1999).

[216] This consideration should be added to the problem with understanding what type of relative value (which should be non-teleological), the deontological reasons respond to.

[217] I am not saying that the adoption of goals is fully optional, as if we willingly and deliberately choose a goal and decide to follow it. I just want to stress that we are able to leave a goal or a project behind, if after reflection we see it is not worthwhile, and only in this weak sense I find goals and projects optional.

The divergence of value argument by itself does not strengthen the position of the agent-relativist theorist: that is, unless the agent-relativist is in a better position to explain the divergence of value as a source of agent-relative reasons in terms other than the presence of goals, projects, etc. Such a master argument is not what I offer in this chapter. Instead, I focus on the presence of goals, projects, commitments, decisions, etc., which is a common both for the neutralist and for the relativist theorist, explanation for the divergence of value. I confine my modest task here to challenging the neutralist theorist from within that shared position.[218]

3.2. Divergence of Value via Agent-Relativity?

The neutralist holds that the divergence between the impersonal value and the personal value of an option (to the agent) does not establish the existence of agent-relative reasons. It is not necessarily the case that the divergence gives rise to a different type of reasons - *agent-relative reasons.* It can as well be the case, that the value of the option gives (1) neutral reasons for action to all possible agents. And it may give (2) *stronger, even additional,* but nevertheless still *neutral reasons* to an agent, whose goals make that option more valuable to him, than if it were neutrally considered (if not taken to form part of anyone's goals).

Consider the agent-neutral explanation of the above divergence as to why it is permissible for me to write the current text. It is, first, that I have a neutral reason to write the text (it is a worthwhile activity per se and hence any agent in a position to perform such action, has a neutral action-reason to perform it). And, second, given my goal (which is also worthwhile), this reason has a greater force, capable of overriding the

---

Constraints are not so optional– irrespectively of whether we consider them valuable or not, we are under duty to follow them and it is not up to us to leave them behind.

[218] This distinguishes my approach both from the position accepting that the only reasons are desire-based Williams (1981) and from such 'hybrid' positions that see in affective desires a source of reasons on a par with value-based reasons , Chang (2004). My position may seem close to this latter one. However, I do not believe desire-based reasons have a fundamental normative status. My modest view is that they are normatively significant, and have to be recognised, but are essentially parasitic on value-based reasons. The difference between the two positions on the status of desires might be put in the terms of the distinction Parfit (1997) introduces between a normative fact and a fact of normative significance: the normative (meta)facts determine which other facts are of normative significance, so the former are fundamental. Desires, then, might not themselves be normative facts, might not be fundamental, though they could still be facts of normative significance for how agents ought to act. They could, accordingly, provide them with reasons for action.

countervailing and otherwise stronger neutral reason for somebody else writing a better text.

I see two possible routes for trying to challenge the neutralist theorist's denial that the divergence of value implies the existence of agent-relative reasons for action.

The first is to attack the claim that the presence of goals *does not add* some *different type* of reason to the stock of neutral reasons - that it just adds strength to the already existing neutral ones. The presence of a goal, it is claimed, can explain the divergence of the value of the option without at the same time implying the existence of additional to the neutral type reasons. Here it is pertinent to discuss whether a specific distinction between goals (adding strength to the reasons, without adding other kinds of reasons) and desires (presumably not even adding to the strength of the ex ante reasons, because the latter entirely determine the desire itself[219]) can be maintained. I claim that at least in some cases *desires do add to the ex ante reasons, on which they are based, and do add different types of reasons*. These are cases when desires act as tie-breakers.[220] Desires can be tie-breakers in momentous (deciding on a comprehensive goal for one's life) as well as in relatively less important occasions. It might, moreover, be the case that this capacity of desires to add to the reasons of the agent is the best explanation as to how the fact of the presence of goals can make difference to the strength of the reasons of the agent. If this is established, it is clear that the description of this additive capacity of the desires cannot be given in agent-neutral terms. If desires can add to the reasons, that, which they have

[219] Dancy's position is representative for the position of those, like Scanlon (1999) and Raz (1999), who deny that desires can be reasons: "Desires are held for reasons, which they can transmit but to which they cannot add. Therefore a desire for which there is no reason cannot create a reason to do what would subserve it." Dancy (2000: 39)

[220] On the difficulty posed by the case of desires, serving as tie-breakers, and thus being reasons (adding to the neutral reasons), for the position that desires do not add to the reasons, see Dancy (2000: 39). I am not convinced by the way he deflects the objection from desires as tie-breakers. He claims it would not be enough for granting desires normative status, to establish that they can be reasons in tie-breaking situations. This is so, because this could easily be done by a reason rather than by a desire, where that reason is left till the moment a tie has been reached, and is introduced only to decide the case. Dancy's solution is not satisfactory: the desires, which were meant to solve the tie, once the ex ante reasons on both sides were evened, cannot be counted among the ex ante reasons themselves. I take Dancy's suggestion to be that the tie could be achieved, by substituting one of the ex ante reasons with a desire, so that the substituted ex ante reason can be left aside (spared) and used later for tipping the balance after a tie has been reached. It seems clear that we have a genuine case of tie, only when a*ll* the *reasons* for and against an action are evenly balanced. Only in this case leaving the decision to the prevalence of desires is warranted. Only in such cases desires could be granted normative status. The cunning move of keeping a reason (by substituting it with a desire) till the last moment, and adding it to decide the case, does not succeed – if not all the ex ante reasons were balanced, there is no tie to begin with.

added, cannot be agent-neutral, since it depends on the presence of a certain desire in the agent alone.

The second route would be to show that the neutralist explanation of the divergence of value cannot explain the case of deontological constraints (the cases where the relative value (or dis-value) of an option for an agent gives rise to deontological reasons). In this case (i.e., if the neutralist is at all willing to recognise that there are valid considerations, best described as deontological constraints) the relative value (dis-value) of the option would ask for a different type of reasons, not simply for adding strength to neutral reasons.

This second route may fail to establish the validity of agent-relative reasons, if it is the case that the neutralist theorist does accept a non-teleological, or a mixed account of value (there are two orders of value with differing requirments). It is not clear to me to what extent the neutralist is committed to accepting a teleological account of value alone. It is often claimed[221] that only a consequentialist (broadly teleological) moral theory can account for the place neutral value (agent-neutral reasons) has in our moral reasoning. The charge against the agent-relativist accounts is that they fail to universalise judgements of rightness. But I see no decisive argument why a deontologist cannot also accept only neutral values, if these values are *not teleologically* interpreted. The argument from the failure of universalising the judgements of rightness might not apply to such a position. A deontologist might agree that it is wrong for A to kill in the given circumstances. He may proceed to universalise that judgement by agreeing that it would be wrong for B to kill in the same circumstances as well. He need not concede, however, that the rightness of this judgement implies any obligation C might have to prevent both A and B from killing by C himself killing (one killing occurs versus two prevented killings). This would be implied only by a teleological interpretation of neutral value, with a maximizing strategy of promoting it. The value (dis-value), a deontologist might point out, is only to be respected (through an action/omission), and need not be promoted.[222]

[221] Philip Pettit's defence of consequentialism (in Pettit 1997) is a case in point.
[222] The distinction between promoting and respecting value is owed to Scanlon (1999).

So, the agent-neutral theorist may accept two types of value: a teleological, giving rise to neutral outcome-reasons, and a non-teleological, giving rise to action-reasons, not demanding maximising the occurrence of the required by this value action (or omission).[223] So, he may well explain the divergence of value in the case of deontological constraints, without introducing relativity of value to the agent, depending on the presence of contingent goals, projects, etc (anyway irrelevant in the case of deontological constraints). The two sorts of value above would ask for two types of neutral reasons: possibly, these could be the already mentioned action-reasons and outcome-reasons.

There might be problems with this picture of two normative orders (two types of value, and corresponding to them two types of neutral reasons). Indeed, the difficulties posed by (1) the priority problem[224]and (2) the requirement of establishing that the two orders are incommensurable, are two of the most obvious ones.[225]

These problems notwithstanding, I do not have a conclusive argument against the possibility of coexistence of two types of value,[226] as sources of two types of neutral

[223] This could be an absolute ban on actively performing actions of killing, lying etc, irrespective of whether by following this directive, the agent fails to prevent performing of such actions by others. The difference is between action that should not be done, no matter what, and minimising the occurrence of such actions (which would be the teleological interpretation).

[224] The priority problem: which type of value should be given priority in cases of direct conflict or a tie in determining the overall value of an option. It is necessary to establish the overall value, because it is supposed to guide the agent in his actions. If one allows for specifying a priority relationship between the two orders (a lexicographic ordering), in order to solve the problem of tie and conflict, one could not possibly claim that the two orders are equally important (and I take the possibility of making this claim to be the main motivation for introducing two normative orders)

[225] The necessity of recognising that the two orders are incommensurable is required if the two orders are not to collapse into a single order. The trouble is that this requirement of incommensurability may be compromised, if the priority solution to the problems of ties and conflicts is not acceptable. Since the overall value is to be determined, if no priority relation is established (the two orders are deemed equally important), this could only be done by comparing the strength of each of the values in this option. But this means that we have a case of commensurability of the values, and accordingly, not two orders of value.

[226] The two normative orders account can be popular among theorists with agent-relativist affinities as well. Scanlon, for example, suggests that "being valuable" cannot always and exclusively be translated into "to be promoted": there are more ways of being valuable than the teleological account of value suggests. Scanlon does not put his position in terms of introducing two normative orders - he supplements teleological value with non-teleological one, without splitting the two as alternative accounts, or as co-existing normative orders. However, he is explicit concerning the priority issue - he admits that there are certain agent-relative considerations (to do whatever is involved in being a good friend) which make certain teleological neutral considerations (to promote the occurrence of friendships) ineligible as reasons, Scanlon (1999: 88-90). His position comes close to the "two normative orders" approach, and might accordingly be criticised along the above lines. The inevitable bias in the way the priority issue is resolved, depending on the neutralist or relativist affinities of the respective theorist, renders the two normative orders solution somewhat incoherent.

reasons. Let me therefore concentrate on the *first route* of arguing for agent-relativity: it promises to establish the plausibility and the validity of agent-relative reasons for action.

3.2.1. Divergence of Value via Presence of Goals

The neutralist is ready to admit that the presence of goals or projects may give additional strength to the otherwise neutral reasons for an action and thus tip the balance in favour of performing it. This means that the presence of a goal adds something to the reasons of the agent for performing the act: it makes a difference. I will concentrate here on exploring the implications of that difference.

Let me note, first, that the neutralist theorist is committed to the view that desires are not neutral reasons for action[227] and since she does not recognise the existence of any other than neutral reasons, she must deny that desires constitute reasons at all. Secondly, the neutralist needs to establish that there is a clear distinction between goals, projects, on the one hand, and desires, on the other. This is necessary, because the neutralist claims that desires cannot add anything to the reasons of the agent (they not only cannot make an invaluable option valuable, but even stronger, they do not add anything to the value of an otherwise valuable option), while goals do add – they make a difference to these reasons.

3.2.2. Goals Only or Desires as Well?

There is a problem for the neutralist, however, with maintaining the distinction between goals adding and desires never adding reasons for action. To see this, let us inquire into what the neutralist precisely means when admitting that the presence of goals does add (makes difference) to the reasons of the agent. Two interpretations are possible. A weaker one - the presence of goals adds some strength to the otherwise already present reasons (the latter are presumably neutral: do not depend on the goals, much less on the desires of the agent). And a stronger - the presence of a goal gives a new reason, not simply

[227] This position is accepted by the non-neutralists as well, though they accept the existence of other reasons along with the neutral ones. Thus Nagel (1986: 167) distinguishes neutral from relative reasons (of autonomy, in particular) precisely on ground of presence or lack of desire as a determinant of the value of the action. Note that Scanlon (1999) denies that desires are reasons for action (and thus sides with the neutralist theorist), nevertheless recognises the existence of relative reasons. It is not entirely clear whether he recognises only deontological reasons (consistent with denying that desires are reasons), or he accepts reasons of autonomy as well. Only an agent-relativist, eager to defend agent-*relative* reasons of *autonomy specifically*, needs to recognise that desires can, in principle, constitute reasons for action.

strengthening the already pre-existing reasons. The stronger better fits Raz's discussion: "He has a reason … that he did not have before. " Raz (1999: 64), and also:

> "The emerging picture is of interplay between impersonal, i.e. choice-independent reasons which guide the choice, which then itself changes the balance of reasons and determines the contours of that person's well-being by creating new reasons which were not there before. This interplay of independent value and the self-creation of value by one's actions and one's past provides the clue to the role of the will in practical reasoning. Previously I have claimed that wanting something is not a reason for doing it….It is, …part of a valid reason for action, once the initial commitment has been made." Raz (1986: 389)

If I am right to urge accepting this stronger interpretation, it would be possible to establish that desires can acquire the status of reason. At the same time they would be different from the neutral type of reasons: they are agent-relative.

To see this, let us move one step back and ask how one adopts a goal, a project, or makes a commitment. Even if there is no deliberate and deliberative process of "adopting" a goal, one thing seems clear: one in principle adopts a goal, gets committed to a cause or a person, develops a project, *for a reason*. Nevertheless, it is agreed that there are cases where there are equally good reasons (or the reasons are incommensurate – it does not matter which of these two cases takes place) to adopt alternative and at the same time incompatible comprehensive goals. In that case, if one makes decision to adopt one rather than the other goal, one surely makes it for reasons, at least in one sense. But since the balance of reasons does not uniquely determine which of the alternative is to be chosen, the fact of the presence of a stronger desire for one of the options *actually decides* the issue. This is captured in saying that desires play the role of tie-breakers: though the decision is taken for reasons, why this decision rather than the equally supported by reasons decision is taken may be accounted for by the presence of (stronger) desire for the decision actually taken. It is *because* of the desire though not solely because of it.

Now, the neutralist would try to make a case against such a desire-as-reason position, by insisting that the "because" above has only explanatory force: can explain what was the motivation of the agent for choosing the actually chosen alternative. It has nothing to do with the "normative" issue - the main concern of the defender of the desires-as-reasons position. This rejoinder, however, begs the question: it is just a restatement of the claim

that desires never add to the stock of the reasons for action. In the case above, accordingly, desires should not be seen as adding to the reasons for adopting the goal.
My response is that if this were so, if the presence of desires played solely an explanatory role, singling out simply the motivation that led the agent to make the decision he ended up with, this would leave unexplained how the goal, which was indeed adopted partly *because* of the presence of desire, once adopted, could add to the *normative reasons* of the agent. It has been already agreed on all sides, that the *value* of an option *is enhanced* by the fact that it is part of a comprehensive goal, thus adding further normative reasons for the option.
To more clearly see where is the problem, we should consider what gives a goal its normative status, allowing it to add new reasons to the reasons that existed before.

This normative status should come, at least partly, from the reasons for adopting it in the first place. If desires cannot add to the normative force of the pre-existing reasons, however, these pre-existing reasons will *exhaust* whatever counts as a reason in the goal. But notice: it was already admitted, that the presence of a goal *adds* to the pre-existing reasons for adopting it. Otherwise the following problem arises: whenever there are equally weighty (or incommensurable) reasons for another goal, one will face at T 2 the same problem as in T1, in deciding which action to follow: the one, favoured by the already adopted at T1 goal, or rather the alternative action, supported by the ex ante equally reasonable alternative goal. If the decision at stage T2 is to be based on the pre-existing reasons alone (presumably exhausting, as the argument goes, whatever normative reasons there are for following the adopted goal), one will once again need to ground one's decision on the greater desire. Thus a desire will decide that case as well (the fact that the agent will most probably have greater desire to further the already adopted goal is beside the point here). Consider the even more interesting case when the eligibility of an action, which would further an adopted goal, cannot be maintained in a situation, where the initial balance of reasons has changed in favour of the ex ante equally valuable but at the time T1 dis-favoured alternative. The decision to stick to the adopted goal in this latter case could not be entirely defended on rational grounds, unless the

reasons in favour of the adopted goal are *not exhausted* by and go beyond the reasons for its adoption.

The problem with a desire-based solution to the decision-problem at this later stage T2 – is that the distinction between goals and desires may collapse. Even in the presence of adopted goals, which presumably are to decide the issue, what has to be done is still being decided on the basis of a prevalence of desires. If this is unpalatable, as I believe it is for the agent-neutralist as well, one needs to claim that the adoption of a goal *adds to the reasons* that supported its adoption. The *normativity* of the adopted goal does not come *entirely* from the reasons that supported its adoption in the first place. What is missing from this picture of goals adding value to the neutral value of an option, is precisely an explanation for this *additional normative* force, which should presumably not rest on smuggling back the desires.

Thus one has to refute the presupposition that the normative status of the goal not only stems but is exhausted by the reasons for adopting the goal. One has to admit, that the additional reasons in favour of a goal-promoting option stem from something other than the reasons for adopting the goal in the first place. But if it was really the case that the desires did not add anything to the reasons for adopting the goal in the first place, what can explain the difference the goal makes to the value of the action, thus adding to the reasons for it?

One may attempt an explanation in terms of the time and emotions already invested, which add to the reasons for continuing to stick to the chosen comprehensive goal rather than abandoning it in favour of the ex ante equally eligible alternative. However, if there was any additional (to the reasons, which were in the cases we are concerned with equally balanced) justification for choosing the actually chosen goal rather than the alternative, as well as for investing time and emotions in pursuing it, it was that the agent preferred it more at the time. It is unlikely, and unreasonable to invest time and emotions in the alternative one disfavours. Though it is not unreasonable to stick to an adopted ex ante equally valuable goal, once one has started to disfavour it, if one does stick to it, it could hardly be simply because one has invested time in pursuing it. Neither time nor emotions invested in what is ex post believed a worthless goal, do justify adhering to it. If adherence to a disfavoured (though not believed worthless!) goal is warranted, it is at

least partly because of the added value, grown *out of the initial desire* that tipped the balance in favour of the chosen goal in the first place. Thus we once again end up with desires as a source of the additional value (though they need not exhaust it – time and emotions may play their role as well), attached to the adopted goals of the agent.

My conclusion is that the distinction as defined by neutralists between goals (adding to the value of the option), and desires (*never* adding to the value of the option, and thus never adding to the stock of reasons as well) cannot be maintained. At least in some cases, desires do add to the value of an option. It is either the case that the goals do not add to the value (but then the observed divergence of value is incomprehensible), or that the desires are the source, that allows the goals to add to the value of the option – and thus desires themselves indirectly add to the value of the option. It is clear that if that conclusion is accepted, the reasons that are constituted by this desire-based added value, cannot be agent-neutral: they should depend on the desires of the agent himself, and thus should be agent-relative.

4. The Limits of the Argument. Some Objections Considered

Before drawing the implications of the agent-relative reasons argument for Raz's Service conception, let me concede two points. I believe these concessions need not be lethal for my argument. First, the explanation of the divergence of value in terms of agent-relative reasons as dependent on the presence of desires does not preclude the possibility of having agent-relative reasons, which stem from the agents' goals, without even indirectly resulting from the agents' explicit desires. The value of options to the agents may diverge from their impersonal value due to the agents' goals even though the agents never had a moment of explicitly deciding to adopt those goals, neither on the basis of reasons nor on that of desires. The emphasis on desires I put in maintaining that the explanation of the divergence of value requires the recognition of agent-relative reasons was due to the fact that desires seem naturally to support agent-relative reasons. It is not critical for my project that the agent-relative reasons are necessarily explained in terms of desires. Goal-based agent-relative reasons for action would do as well, once it is recognised that the increment of value the presence of goals adds to the ex ante value of an option provides reasons of an altogether different kind - agent-relative reason for action.

Secondly, I should stress the parasitic nature of the desire-based value, which constitutes the agent-relative reasons. I tried to establish that the increment of value, that adds to the strength of reasons, normally (though not always) comes from the fact that certain goal is more strongly desired than the equally (or incommensurably) rationally eligible alternatives. However, this does not mean that the whole value of the option is exhausted by its value for the agent, who has adopted (because he desired it more) certain goal. Part of the value of the option will give an agent-neutral reason for action to all agents, our agent included. The additional to that value, value for the agent, who has as a matter of fact adopted a goal, which will be furthered by the prospective action, will constitute a further, agent-relative reason for that agent alone. This relative reason will not always be conclusive (and the neutral reason, which all other agents have, will not necessarily always be defeated either): the agent might be rationally required to act on a stronger neutral reason, which will render the action, for which he has a relative reason, ineligible. This explains how some other agents may have a neutral reason for action, which would promote a goal of an agent, (given that goal is valuable), as well as how the relative reason of that agent may be outweighed by the disproportionately greater neutral value of an option (which may not promote a goal of that agent).

A further problem with my construction is that the extent of the divergence of value, which determines the strength of the relative reason, may be taken to depend entirely on the desires of the agent. It might be pointed out,[228] that this would fail to explain why we take the success of one's pursuits to matter: it would be implausible to claim that we care about that success, simply because we desire those pursuits, goals, etc. Even if we happen to get 'colder' to them, we will still have a reason to succeed in them which would be different than the neutral reason: we would have failed, and this failure would add to the (dis)value for us of that pursuit. Thus desires cannot be the whole story.

As in the above rider, let me again point out: maintaining that agent-relative reasons are necessarily desire-based is not critical for the success of my project. It might be true, that the desires are just an enabling condition, triggering the adoption of a goal, which may grow into a comprehensive goal, giving agent-relative reasons for action. What is important for my purposes here, is that the fact of divergence between the neutral value

[228] This objection was raised to me by Professor Raz.

of an option, and the value of that option to an agent, no matter whether it is explained by the presence of desires, the desirability of succeeding in one's pursuits, the fact of a choice, the importance of deciding on a course of action, or something else, securely establishes the existence of agent-relative reasons.

There is another concern with my claim here: it is too strong.[229] If it is true that the divergence of value establishes the validity of agent-relative reasons, and these latter support a liberal-democratic political arrangement, then this would be a universalist defense of liberal democracy. All societies, whose political arrangements were different, would have failed to pay due respect to the persons, because they did not provide for them the opportunity to act on their agent-relative reasons. This would mean that I condemn all past political arrangements.

My response to this objection may seem somewhat arrogant. My claim is that even though the past political arrangements were not to be condemned for failing to pay due respect to persons by recognising their agent-relative reasons for action, they were still wrong not to do so.

A more sophisticated, though again rather controversial response, might appeal to ideas, developed by Raz himself. [230] Start from the position that there might be no timeless moral verities. Morality may, nevertheless, still "continuously and endlessly develop toward unchanging moral principles." Though these unchanging moral principles might not have been valid at all times, since they were (for different reasons) 'beyond people's grasp,' those subsequently valid principles may vindicate the then-valid principles. Thus, there might be a universal, unchanging moral principle, stating that since people have agent-relative reasons, they should be allowed to act on them directly. At the same time it might be recognised that past political arrangements need not have been wrong to not have followed this moral principle in all times, since it might not have been valid then. This argument may support the position that liberal democracy is a universally justified form of political organisation, at least partly because it relies on and realises such universal moral principles.

[229] I owe this objection again to Professor Raz.
[230] Raz (1994: 156-157).

The difference between the arrogant and the sophisticated position is that on the first, people are guilty of failing to respond to a valid moral principle, though they are not to be blamed since they might not have had "access" to it. The second, more radically, postulates "moral change": the then valid principles did not require recognition of agent-relative reasons for action (so people were not even wrong, and not simply not blameworthy, not to recognise them,) though the fact that there are such reasons may be a universal, if not timelessly valid principle. Clearly, the concept of universally valid moral principles that are *not timeless*, is not uncontroversial. I agree, though, that there is something odd in the first position as well. Saying that though people were wrong but are not to be blamed on account of their incapacity to grasp the universally valid principles, may be no less controversial.

A less controversial (not based on controversial moral ontology), but not necessarily sound reaction to a similar argument - autonomy is valuable only because of the conditions in modern Western societies that require individual choice, and is not universally valuable, is owed to Waldron.[231] He asks why if autonomy is only valuable because of the conditions in such types of society, it is state's duty to maintain those conditions (this latter claim is one of Raz's main theses[232]). If the value of autonomy was not extending beyond its value in those conditions, there is little ground for this duty. The analogy Waldron draws here is revealing: if Raz is right about the value of autonomy, why is not the same with the virtue of justice. Both make sense only under certain conditions – one under conditions of modern liberal, individualist societies, the other – under conditions of scarcity. If the virtue of justice does not justify preserving the condition of scarcity, so is with autonomy. If there is an important duty to maintain its conditions, however, this indicates that the value of autonomy is not indexed to a type of society.

Raz's response [233] is that the state has this duty of supporting the conditions of autonomy only in the circumstances of normal politics (i.e. when no radical change in the character of society is involved) and does not extend beyond that. If so, Waldron's argument fails.

[231] Waldron (1989: 1122).
[232] In Raz (1986) and Raz (1994).
[233] Raz (1989: 1228-9)

So, we are left with the arrogant and the sophisticates responses to the "universality" argument to choose from.

A further, general objection to my account of agent-relative reasons might point out that the cases I have used to push forward my case, are so peculiar and exceptional, that they cannot vindicate the general recognition for the existence of two types of reasons. Thus Raz, for example, recognises that in these cases desires function as reasons, but nevertheless does not accept them as normal reasons. [234] The agent-relative reasons in these cases are too peculiar.

But if incommensurability (one of the two possible sources for accepting desires as reasons) of the value of the options, and hence of the reasons for actions is wide-spread,[235] the desire-based agent-relative reasons would not be too rear, even if peculiar. They might not be exceptional. Moreover, besides cases of incommensurability, cases of equally weighty options could also trigger my argument: so, even if some find the idea of incommensurability of options impossible, they could still recognise the validity of my arguments. If still some find the concept of a desire-based reason for action peculiar, I have conceded the possibility of having non-desire-based agent-relative reasons, for which neither the presence of incommensurability, nor that of equally weighty options is critical.

If the above argument for the implausibility of the neutralist explanation of the divergence of the neutral value of an option and the value of the option to the agent is sound, the next problem I need to address is whether there are some additional sources of discontent with the agent-relative reasons. If there are such sources of legitimate discontent, one has two routes open. One may try to find an alternative to the criticised above neutralist explanation of the divergence of value, which will need to be still neutralist in character. Alternatively, one may need to claim that there is actually no divergence of value: what seems to us divergence of two types of objective value, might as well be a possibly justified (because it is due to the natural partiality of our personal perspectives) distortion of the only true neutral value.

[234] See his "Incommensurability and Agency", in Raz (1999: 62).
[235] For this claim, see Raz (1999: 66).

A possible source of discontent with the agent-relative reasons, added to their alleged peculiarity, and possibly one of the springs thereof, is that they seem not to fit some of the plausible requirements for an account of reasons for action. The two requirements[236] are: first, that the evaluative properties of the options (the value of the action) serve both evaluative (judging the action) and action-guiding functions, and second, that the domain of values and reasons is intelligible. I take the second to be a direct consequence of the first: if the value of the action is to guide the agents in acting, this value should at least in principle be intelligible (even if not always explicitly explicable) to those agents.

Why would agent-relative reasons not fit these requirements? For one, the desires we took sometimes to constitute agent-relative reasons, are only to an extent intelligible: they are intelligible to the extent they are held for reasons, but once the reasons are evened, it is not intelligible why one desires one option rather than the equally valuable alternative. And if these only partially intelligible desires are allowed to enter the domain of values and reasons, this domain would loose (or lessen) its intelligibility. These problems with the intelligibility of desires are connected with the requirement that values should serve both evaluative and guiding functions. If one desires more one rather than the other of two equally plausible competing options for no particular reason (and it is unintelligible even to him why he desires more one rather than the other), it is not clear how his choice of one of the alternatives is at all guided.[237] What is in principle unintelligible cannot normatively guide action. The value of the alternative options is intelligible, and can guide the action, but in the case of a tie, it is precisely the unintelligible desire that tips the balance in favour of the chosen alternative, and thus actually serves the guiding function. Thus the two functions of value come apart: value serves only the evaluative function: it testifies, that the alternative options are eligible; however, the guiding function is served by desires. And desires by definition cannot serve this function, because they are unintelligible.

To deal with this objection, one may try once again to proceed from the case of goals as reasons (which case presumably satisfies the two requirements for an account of reasons) towards the case of desires (which I take to underlie the choice of goals at least on the

[236] These requirements are advanced by Raz in "The Truth in Particularism", in Raz (1999: 219-220).

[237] The explanation of action in terms of reasons seems to require that the agent be taken to have acted *for* particular reasons, or to have been *guided by* those particular reasons.

occasions of incommensurable or equally valuable goals), and then challenge the distinction between goals and desires. Thus the compatibility of goals as reasons with the structural requirements for an account of reasons, could be extended to cover the compatibility of desires-as-reasons with those requirements.

However, it might be that this move backfires: instead of expanding the field of considerations that count as reason to cover desires in the cases discussed, can have an adverse effect on the recognition even of goals as reasons. If it is true that the added value of the option, when goals are present, is attributable (in the above cases) to the fact of the presence of desires (which are unintelligible, and cannot guide action), then so much worse for the goals: they will turn out in fact *not* to add to the value of the option. But, then, the critic would need to advance an explanation in neutralist terms (not based on the presence of goals, projects, etc.) for the observed divergence of value, and it is not clear whether it would be successful.

Alternatively, it is possible to go one step further, and deny that there is a divergence of value on occasions, where desires tip the balance, and goals are seen as determined partly by such desires. It is counter-intuitive to claim that precisely in cases of equal or incommensurable value, the value for the agent of the option the agent has actually chosen as his goal, is not greater than its neutral value. If there is divergence of value at all, it will certainly apply in these cases in the first place.

These cases have the advantage as well of providing us with a relatively clear *counterfactual test* for establishing what exactly has been added to the value of the option, once adopted as a goal. The move of denying the divergence of value in such cases, would be unpalatable for those embracing the common sense morality (with its recognition of the divergence of value in these cases), and aim to provide an account of reasons for action, not radically departing from it.

It will be noted, of course, that my strategy in replying to the objection from the requirements for an account of reasons, was to retreat to a point, where the opponents would feel uncomfortable with the results of their attack. This strategy did not (nor was it capable) establish, however, that the desire-based agent-relative reasons can meet the above requirements. If it turns out that the consequences of denying the validity of such reasons are truly unpalatable, it might be worth considering the possibility of relaxing to

an extent these requirements, which would allow for accommodating within the admissible accounts of reasons, agent-relative reasons as well.

5. A Reason-based Account of Autonomy

I have insisted that the central case for agent-relative reasons in general is the divergence between the neutral value of an option and its value to the agent. In the discussion above I have concentrated on the agent-relative reasons of autonomy, leaving aside the difficult issue of deontological reasons. May be the divergence of value can explain their case as well. But my main concern was the contested though common ground between the neutralists and the relativists: which is the better explanation - in terms of relative or neutral reasons, of the divergence of value, if it is agreed that it can have its source in the goals of the agents. I have tried to show that the neutralist theorists, who deny the validity of agent-relative reasons, while accepting this divergence of value, cannot give a coherent account of this divergence. The distinction between desires and goals,[238] on which their neutralist explanation of the divergence of value relies, seems difficult to maintain in the form required for their task.

I think the appeal of the divergence of value thesis, is in that it helps spell out the idea, underlying the ideal of personal autonomy: we should be free to be authors of our own lives, precisely because there is a divergence between the neutral and the relative value of one's options to the agent. With the danger of making a hasty generalisation, it may well be that the recognition of the divergence of value underlies the whole individualist liberal tradition.

It seems, nevertheless, the neutralist has a strong case, apart from the considerations dealing with the divergence of value, against the agent-relativist. He might insist that he offers a reason-based account of autonomy, and consequently, of liberalism, which is distinct and superior to the will-based (in the sense of thick will, desire-based) one, allegedly defended by his relativist opponents. The neutralist claims that since autonomy has a value to the extent it follows right reason[239] (otherwise it is blind and with no

[238] It should be clear that I am not denying that there is a distinction between them – I only deny that the main distinction between them is that one does, while the other does not at all add to the neutral pre-existing reasons for action.

[239] See Raz (1986: 318): "Autonomy is valuable only if exercised in pursuit of the good."

value), his reason-based account of autonomy better protects this value of autonomy by not allowing desires to play any role in it - precisely because desires are of no relevance for that value. If one agrees with the neutralist on the point about the value of autonomy, one is to restrict, this argument runs, the relevance of desires to the dis-favored will-based account of autonomy. It is clear why the will-based account of autonomy should be dis-favored: it makes the very implausible claim, that the *value of autonomy* is in the satisfaction of the *desires* (preferences) of the agents, because desires alone provide the agent with reasons for action. This position makes the relativist uncomfortable in the same way the neutralist was made uncomfortable in the previous section: because he cannot explain the fact of the divergence of value. If all the value for the persons is exhausted by the desires of the agents, there is no divergence between the neutral and the relative value of the option: it is simply denied there is such a thing as neutral value of relevance for persons.

Now, nothing in my discussion above suggested that I would like to defend this strong will-based account of autonomy. After all, I try to show the recognition of agent-relative reasons is necessary for a reason-based account of liberal democracy, presumably itself favoring a reason-based account of autonomy. I need not deny the value of autonomy consists in following right reasons. I simply try to establish that *desires* can, in certain situations, be such *right reasons*. I am not defending the strong claim, that whatever value an option has, it has it because it is desired. On the contrary: I simply claim that desires can provide us with additional reasons to favor an otherwise already valuable *per se* option, thus singling it out among otherwise equally eligible alternative options.

What I find disconcerting in the neutralist defense of a reason-based account of autonomy, is precisely this denial that desires can provide us with valid agent-relative reasons for action. This position threatens to undermine the only explanation of the divergence of value the neutralist can provide – in terms of the goals of the agent. If it is true that the recognition of the divergence of value is what supports the ideal of personal autonomy as morally relevant in the first place, the failure to explain this divergence of value disadvantages the neutralist position. The reason-based account of autonomy is not threatened by this challenge to the neutralist position: to the extent that not brute desires,

but reasons (agent-neutral and agent-relative alike) are taken to support the value of the autonomous choice of options, one works within a reason-based account of autonomy.

6. Conclusion: The Autonomy Condition and the Normative Supremacy Claim

This discussion is relevant for my thesis, let me stress it, since I take the reason-based account of agent-relative autonomy, and not the will-based one, to be a strong support for Raz's autonomy condition. In evaluating whether an exercise of authority is justified, one is first to show that it is more important to decide an issue correctly rather decide it on one's own, and only then run the legitimacy test of NJT. Some of our autonomously adopted goals require that we are left alone to decide how best to promote them, even if this would bring lower, by some external standard, level of conformity to those goals. The point is that the value for us of pursuing our goals on our own, even at the expense of getting sub-optimal results, is out of proportion with the value of achieving these goals, when aided. This is not the position defended (unsuccessfully, as I claimed in the previous chapter) by the strong endorsement constraint thesis: that an option has no value whatsoever unless endorsed by the person.

The claim defended here is modest – it often is more important, and brings more value, to act on one's own goals directly, rather than pursue them indirectly by following authority instead. The latter does not entirely deprive them of value - it just significantly reduces it. However, even this modest result seems to substantially limit the scope of the NJT.

More importantly, the result here shows why the claim authority necessarily makes of having comprehensive supremacy over all other normative domains, cannot be plausibly made. If the presence of agent-relative reasons of autonomy necessarily limits the justified exercise of authority, it is those reasons, and not authority that determine this limitation. It may be this limitation - that authority should respects such reasons and thus should necessarily refrain from making the utterly implausible claim to normative supremacy over all other normative domains, that best characterises the liberal-democratic form of political authority.

This challenges Raz's conception of political and legal authority and their justification in a liberal-democratic political order. First, the autonomy condition shows why political

authority cannot make a bona fide claim to supremacy over all other normative domains. Even less can *liberal-democratic* authority, on which the restrictions of the autonomy condition are to be taken even more seriously, make such a bona fide claim. However, if making this claim is an essential feature of political authority, on Raz's model, this not only shows its internal problems as an adequate conception of this authority in general, but also threatens its applicability to the liberal-democratic one specifically. This is so, since it is likely a defining feature of this type of political authority in particular, that it refrains from making such overboard claims to supremacy over all other normative domains. Moreover, making such claims cannot be a central feature of the concept of political authority, since the case of the liberal-democratic type of political authority falls within the core of the concept of political authority.

Let me now go back to the main hypothesis about the connection between a reason-based justification for liberal democracy, and the recognition of agent-relative reasons for action.

I have claimed that if there is justification for the adoption of certain political principles, it is because by following them the state authority is more likely to issue directives, which would provide its subjects with valid protected reasons for action (helping them better to conform to the reasons that directly apply to them). If I am right in insisting that agents have agent-relative reasons along with their agent-neutral ones, it might as well be the case that the guiding political principles need be of a liberal-democratic character. The subjects' agent-relative reasons of autonomy and partiality ask for a constitutional protection of their individual rights, and, in addition, their reasons of autonomy may require the adoption of certain democratic procedures of decision-making. The latter procedures should, however, be constrained by the constitutional protection of individual rights, if they are to result in authoritative directives, capable of providing the subjects with valid protected reasons for action. Only in this way, it seems, would authority meet the requirements of the Service conception of legitimate authority: to serve its subjects by helping them better conform to their reasons. However, this service is severely limited by the requirement of meeting the autonomy condition – the service depends on whether subject's reasons themselves allow for trying to bring improved conformity to them by

acting on some other reasons. The presence of agent-relative reasons for action seems to boost the case for this requirement being a rather strong one: it is quite often the case that acting on one's own is ways more important and brings more value for the person concerned than acting correctly, led by authority's directives. More than that: it is more important that it is the subject himself, who decides whether the issue is such to require deciding oneself rather than deciding correctly. This is the broader interpretation of the autonomy condition, urged by Green (1989: 811), which seems resisted by Raz (1989: 1180-3). If the broader interpretation of the autonomy condition is more in line with the agent-relative reasons thesis defended here, as I think it is (these reasons are goal-dependent – deciding which goals to adopt is up to the individual, not the authority), this shows that the constraints on the legitimate exercise of authority are externally – by the subjects themselves (as required by this interpretation), and not internally determined - by the authority itself.

Further, recall the discussion of NJT in the second chapter and the problems I identified with it, concerning its capacity to explain in what sense the legitimate authorities, when acting within its legitimate bounds, can turn mere oughts into duties. One of the problems was that an important difference between rational requirements and duties is precisely the independence of the latter from persons' own goals. This consideration may argue for restricting even further the legitimate exercise of authority - only within the bounds of serving directly only the goal-independent, agent-neutral reasons of its subjects. It is clear that if this is so, NJT would not be an adequate legitimacy test. But I will not pursue this at this point - I will say more on NJT as such an adequate test in the concluding part of this thesis.

However, as I already indicated, I believe there is another deep problem with the Service conception, and with the NJT in particular. My concern is that the claim to normative supremacy, taken by Raz to be an essential feature of law's and state' authority, not only cannot be reconciled with a liberal-democratic type of authority, but is inadequate for any type of political authority, since it is in tension with the central normative tenor of the NJT and the Service conception of legitimacy more generally. This conception necessarily tests the legitimacy of law and political authority by moral standards: whether

law and state authority are ultimately morally justified (or required by practical reasons). It is morality, or practical reason, that ultimately justifies authority: but then authority's claim for supremacy over all other normative domains is necessarily false, and thus obviously implausible.

My contention, then, is that making obviously implausible claims cannot be an essential feature of authority.[240] If Raz is right that it is in our concept of political authority, that authority necessarily makes such a claim, then it is the account of justification, on which this claim turns out to be necessarily false, which is inadequate. If rather it is the morality, or practical reason-based account of legitimacy that seems central to our concept of authority instead, then the claims authority is thought to necessarily make should be modified. In either case, something should be given up: either the claim itself, or the type of account of legitimacy that falsifies this claim. I hope to have indicated clearly and persuasively why I believe that the two do not and cannot go well together.

[240] I am not alone in this. The discussion on these issues is expanding fast and is becoming very sophisticated, Himma (2001), Edmundson (2002), Lefkowitz (2004) are but just a few examples.

Part Three

Authority and Instrumental Rationality

In this part of my thesis I move away from the peculiarities of political authority, explaining its distinctness from other types of practical authority, and go back to the fundamental issues, affecting any type of practical authority – the issues around the paradox of rationality involved in obeying authority. I address a specific problem under this general heading, connected with Joseph Raz's account of authority. For Raz, authority, by issuing authoritative directives, purports to give its subjects protected reasons for action. When the authority is legitimate, it does indeed provide subjects with such protected reasons. The legitimacy of authority is mainly established on instrumental grounds. The question I address is whether it is individually rational to decide to follow an instrumentally justified authority, if to follow authority means to take its directives as protected reasons for action. This question, it should be noted, is different from, though connected to the widely discussed question whether it can be rational to follow authority (obey authority) on particular occasions. The target of the first chapter of this part is Larry Alexander's suggestion that deciding to follow authority might not be rational. In the course of arguing for this claim, he draws an analogy of the case of authority (and of serious or mandatory rules more generally) with that of Gregory Kavka's Toxin Puzzle[241] [henceforth TP] of instrumental rationality:[242]

> "…[.serious rules] may be like the intention to drink the vile potion tomorrow in Gregory Kavka's Toxin Puzzle." Alexander (1999: 53).

[241] Kavka (1983).

[242] The direct analogy Alexander suggests is between the possibility of having "serious rules" and TP. For Alexander the concept of legal and political authority is "bound up with the existence of serious rules." Thus the case of political authority, if this author is right, may be analogous to the TP as well. Raz (2001a) also takes the existence of rules to be central for legal authority. He claims that authoritative directives - "one subspecies of mandatory rules" Raz (1990b: 191), promises and more generally commitments, have by their nature one and the same structure: they all are *believed* to provide protected autonomous reasons with exclusionary force to those to whom they apply, and *actually provide such reasons when valid*. I concentrate in this chapter on the case of authoritative directives, since political authority is my main concern.

The question I ask is whether such an analogy could be drawn. Though I conclude that the analogy does not hold, my discussion helps illuminate another problem for the rationality of deciding to follow authority, which I discuss in detail in the second chapter. Is the strategy of always following what one believes to be a legitimate authority, even when one disagrees with its directives on a particular occasion and happens to be right to disagree (since the authority did not get the balance of ex ante reasons right) rational? And if rationality requires allowing room for exceptions, does such a rational strategy have resources for solving the "instability problem": if one is always tempted in cases of disagreement with authority to disregard its directives, and gives in often enough to this temptation, one may end up being worse off by deciding to follow authority than if one always followed one's own judgement only instead. Does not that argue for the irrationality of deciding to follow authority if no stable decision-making strategy is available? It is often assumed that deciding to adopt serious rules (deciding to follow authority) does not present problems of rationality: such problems affect only the subsequent, actual following of those rules, or complying with authority:

> "…according to the naïve compatibilist position, as well as the Constraint and Resolute models [the all main positions], it *can be rational to adopt rules*…The problem for the naïve compatibilist comes when one turns to the rationality of *actually being guided by* such rules." Shapiro and McClennen (1998, 366-367, emphases added)

Is this shared view warranted, or rather, there are problems of rationality already at the stage of deciding to follow authority or adopt rules? My conclusion is that there are indeed such distinct problems. More importantly, I claim that neither Raz's own account, nor any of the recently offered accounts of rational dynamic choice, when applied to this account provide an easy solution to them.

Thus the doubts concerning the rationality advantage of instrumentalist justification for the exercise of authority seem well-grounded. This conclusion reinforces the general critique against Raz's instrumentalist justification of political authority. Recall, the concern is that instrumental justification falls short of what is required if Raz's general account of practical authority is to be a plausible account of political authority as well. A *moral* duty to obey is commonly thought owed to a legitimate political authority acting within the bounds of its jurisdiction, and at several points in my thesis I have argued that

instrumental justification meets special difficulties in grounding such a moral duty. But may be it is necessary to revise (a much-loved by philosophers exercise) our common-sense notion of duty to obey, in order to bring it in line with the philosophically best-supported notion of legitimate authority? This seemed the obvious route to be taken, since the instrumentalist justification promised an indispensable rationality advantage – it offered a solution to the rationality paradox that plagued the inherited theorising on authority. The conclusions reached here - this rationality advantage is suspect, instead of arguing for such revision, should rather prompt a search for other, non-instrumentalist justifications, more in line with our common-sense notion of legitimate authority. In short, going beyond a generally instrumental justification in the case of political authority seems warranted.

Chapter Six

The Rationality of Deciding to Follow Authority: The Toxin Puzzle Analogy

I start this chapter by setting out the structure of the Toxin Puzzle (henceforth TP) case and look at the structural features of authority, in search for analogy between TP and the case of deciding to follow instrumentally justified authority. After dismissing some preliminary objections against the analogy 'TP - rationally deciding to follow authority,' I focus on the main problem for establishing it. The problem is whether the case of deciding to follow authority involves what Kavka calls "autonomous benefits." I show that it does not generally involve such benefits, or even if it does, this is not known to the agent, which is necessary if the case of authority is to be analogous to the TP case. Thus the analogy fails. Nevertheless, though there is no similar to T P puzzle in the case of rationally deciding to follow authority on the Razian account, there is a connected problem with offering a stable rational strategy of following authority. Discussing it is the task of the second chapter in this part of my thesis.

1. Toxin Puzzle and the Instrumental Justification of Authority

Authority, according to Raz, is primarily instrumentally justified. This follows from NJT – the justification for following authority is instrumental: one is normally justified to follow authority when one does better in conforming to one's own ex ante reasons by complying to authoritative commands than he would do by complying to one's own reasons directly.[243] Next, the presence of (legitimate) authority is taken to make *practical difference*[244] to how subjects ought to act, to what reasons they have for action. This is closely connected with the instrumental role of authority. This role presupposes that at least on some occasions authoritative directives require actions that diverge not only with subjects' all things considered *judgement* on the ex ante reasons for action, but with *their*

[243] "Where there are advantages in having authorities…they are always a result of the indirect strategy for conformity with reasons, i.e. maximizing conformity with reasons not by trying to comply with them but by following someone else's judgement about what one should do." Raz (1990b: 195)
[244] I have discussed it in more detail in chapter 2 of my thesis.

ex ante reasons for action themselves as well. The connection to NJT is that only by making practical difference could authority bring the benefit of improved conformity. Next follows the autonomous reasons thesis: the difference-making characteristic of authority, according to Raz, is a function of the presence of "autonomous reasons,"[245] (just another name for the protected reasons discussed in first chapter of this thesis). They include content-independent reasons (CiRs) - typically provided by promising, undertaking commitments, making binding decisions, deciding to follow mandatory rules, authority etc.), deriving their force not from the reasons, justifying deciding to follow authority,[246] etc., but rather by the very fact that such acts have been performed with the intention of creating new, i.e. autonomous reasons for action.[247] This characteristic (the source of their force is not exhausted by the underlying reasons for the promise, the commitment, etc) may partly explain how their presence can make practical difference to the ex ante reasons subjects already have. Autonomous reasons, next, are *"protected"* – the exclusionary reasons (ERs) protect CiRs by excluding all pre-existent, content-dependent reasons within their scope of application, thus making CiRs decisive in resolving issues within their jurisdiction.[248] The protected reasons, however, are neither absolute, nor are they only prima facie.[249] They exclude all considerations within their jurisdiction, and thus are conclusive there, but are not always conclusive all things considered (there might be reasons against the action they require, which are outside this

[245] Raz (2001: 12)

[246] The analogy with promises/commitments holds here too: for Raz they bind even if there were no reasons for giving the promise, committing oneself in the first place, Raz (1986: 388). This binding need not be absolute.

[247] "The Promise keeping principle and the Decision principle are both based on the idea that people should have a way of binding themselves by intentionally creating reasons for action" Raz (1990: 69). The puzzle is "how can it be that people can create reasons just by acting *with the intention* to do so" (Raz 2001: 5, emphasis added). Raz's explanation for the binding force of promises relies on their character of voluntary obligations. "Promises are voluntary obligations not because promising is an intentional action, but because it is the communication of an intention to undertake an obligation, or at any rate to create for oneself a reason for action" Raz (1977: 218).

[248] I am discussing the many problems with the *concepts* CiR and ER in chapter 1.

[249] The duties stemming from the normative practices of promising, committing oneself etc., are sometimes taken to provide new reasons for action, which are *decisive ceteris paribus*, giving rise to *prima facie duties* only; Harman (1978:114), Searle (1978). They might, on the other hand, be taken to provide *conclusive reasons for action* (giving rise to *absolute duties*). Raz takes the middle position: duties have limited absoluteness. They exclude all other considerations within their scope of application, but need not be conclusive "all things considered." Raz explains this by introducing the concept of an autonomous reason with an exclusionary force. For Gans (1992: 21-22) duties also are conclusive reasons only within a limited domain (they are duties with limited absoluteness), but his analysis does not rely on ER concept.

jurisdiction). ERs thus are reasons not to act[250] on those ex ante reasons, which apply directly to the subjects and on which the authority was meant to ground its authoritative directives.

We saw in the first part of this thesis there are many problems with the claim that the autonomous reasons have such *practical* exclusionary force. This characteristic is necessary part of the autonomous reasons, explaining how they can make the required type of practical difference. If found truly problematic, the autonomous reasons explanation of how authority makes such difference may be unsatisfactory. An explanation will be due for why deciding to follow authority, if it implies accepting that authoritative directives provide "protected" reasons for action, is not irrational.

In this chapter I address the problem with one attempt to set the conditions, under which deciding to follow authority is not irrational. Successfully setting such conditions implies that when these are satisfied, the autonomous reasons with their exclusionary element are valid. On Raz's Service conception, let me repeat again, the main condition, establishing the rationality of following authority[251] is, that following authority instrumentally *serves* its subjects by bringing better conformity to their ex ante reasons.[252]

The issue I discuss here is whether there are special problems with such an instrumental justification, when coupled with the main characteristic of authority: that it provides protected reasons with exclusionary force, meant to preempt acting directly on the ex ante reasons. I focus on a particular challenge to this solution to the problem of rationally deciding to follow authority. It is presented by the Toxin Puzzle (henceforth TP) of

---

[250] ERs are reasons *for action*: excluded is *not reasoning* on the merits of acting on the balance of the ex-ante reasons, but only *acting* on this balance.

[251] The same is true for the rationale of adopting mandatory rules, and only in a modified form in the cases of promising and undertaking some specific commitments. Notice that the rationale for taking promises and commitments to impose duties on the promisor is to enable having "voluntary special bonds with other people" Raz (1986:175), which may be of *intrinsic* value. This rationale, then, renders promising, committing *constitutive elements* of the *intrinsic value* of having such special bonds with other people – they thus may not be simply instrumental. I do not address here the plausibility of such non-instrumental justifications.

[252] Two versions of instrumental justification are possible. In the framework of discussions Bratman (1998), Gauthier (1998b), Harman (1998) of the Toxin Puzzle, the instrumental rationality framework implies that the agents' *preferences and (subjective) values* define what is rationally justified. Raz's instrumental Service conception of authority takes the subjects' *ex ante (objective) reasons for action* to define what is rationally justified: whatever brings improved conformity to those reasons. This difference between *subjective preferences* and *objective ex ante reasons* for action will be relevant in discussing Bratman's modified "sophistication" strategy of dynamic choice.

instrumental rationality, showing that the instrumental rationality of an action may conflict with the instrumental rationality of an intention to perform the same action.

Toxin Puzzle: A reward is announced for forming tonight an intention to drink a mild toxin tomorrow afternoon. The reward is given just after midnight tonight (before the intended act of drinking the toxin is to be performed) and getting the reward is conditional only on successfully forming the required intention, not on performing the action for which it is an intention. The question is whether rationality allows one to form such an intention.

On the one hand, it is instrumentally justified to decide (and form an intention) to intake a toxin tomorrow, because thus one expects better to conform with ex-ante reasons (one will get a reward). On the other, there is no reason to act on this decision, subsequent to getting the reward, since there are no further advantages to be gained from performing the act itself, and there are costs involved. However, since one has at the time when one is to form the intention, the correct belief that one will not have a reason to drink the toxin subsequent to getting the reward, one cannot *rationally form that intention*. Instrumental rationality requires something, which instrumental rationality prevents at the same time. This is the puzzle of instrumental rationality, revealed by the TP.

"Toxin Puzzle" in the case of *rationally deciding to follow authority?*

Instrumental rationality justifies deciding to follow authority, because following authority brings improved conformity to reasons. However, following authority involves taking the authoritative directives as giving to the agent protected reasons for action with exclusionary force. If taking those directives as protected reasons with exclusionary force is irrational, instrumental rationality may prevent one from doing what is instrumentally rational: deciding to follow authority. Hence the puzzle.

If there is an analogy between the TP and the instrumental justification for having authority, one of the proposed justifications for following authority may be suspect. But is there such an analogy? This is the first issue I address in this part of my thesis.

2.1. Analogy TP – Instrumental Rationality of Deciding to Follow Authority?

The problem for rationally deciding to follow authority is, roughly, this. I may be instrumentally justified to I decide to follow authority [comply[253] with authoritative reasons] regarding a range of issues, because thus I expect to achieve better conformity to the reasons that apply to me, regarding that range of issues. However, I can see that complying with the authoritative reasons[254] on every occasion is not strictly necessary for achieving better overall conformity to my ex ante reasons. Such compliance is known to be strictly necessary only in the case when authority is perfect. Authorities, even legitimate ones, do make mistakes about the balance of the ex ante reasons, and this fact is known to their subjects. Perfect authorities aside, then, it seems instrumentally justified to refrain from complying with authoritative reasons when they greatly differ from the balance of the ex ante reasons. Thus, there is an apparent analogy with the TP case: the instrumental rationality of the required action contradicts with the instrumental rationality of the decision, on the basis of which it is required. I have an instrumental justification for adopting the rule/for deciding to follow authority, which is not at the same time a justification for following through with the action this decision/rule requires. But is the analogy, nevertheless, only apparent?

2.1. The Structure of TP

Let me first set out the structure of the TP, and see whether the case of the instrumental justification for deciding to follow authority exhibits the same structure. If they have the same structure, the analogy holds.

Toxin Puzzle case:

1. Instrumental rationality framework[255]

2. Presence of an autonomous benefit: there is a benefit to be gained by forming an intention to act, which is *causally independent of*, i.e. autonomous from, actually

[253] The distinction conformity/compliance in Raz (1990b: 182) was discussed in detail in chapter 1.

[254] I concentrate on authoritative (i.e. provided by authority) reasons. They share important features with the reasons provided by mandatory rules, promises, commitments: they all are autonomous, CiR, protected by ERs. Since the justification for having valid promises may differ from the instrumental justification for having mandatory rules and to follow authority, my argument will affect only the latter.

[255] The "standard" view of instrumental rationality is used here: the rationality of an action (in no unanticipated information cases) depends on the agent's evaluative ranking at the time of action of options available then. Two main alternatives to this view were proposed: sophisticated and resolute view, McClennen (1990: 12 – 13). I will define and discuss them later in the body of the text.

performing the intended act. The reason it is impossible to rationally form an intention to act, is the presence of autonomous benefits: these benefits justify only forming the intention to act, but not the execution of that intention.

3. No unanticipated information or unanticipated change of evaluative ranking. This is expressed in Bratman's "linking principle" about the rationality of forming intention: c*eteris paribus*, if one forms a rational intention to A, it is rational for one to A, when the time comes.[256]

At first glance, only the first component is present in the case of the instrumental justification of authority. The main obstacle to drawing the analogy is the requirement of having autonomous benefits. The benefit of doing better by following authority is *not causally independent* from actually following authority's directives. The decision *alone* to follow authority will not confer *by itself* any benefits on the subjects, unless it is coupled with a sufficient degree of compliance with the authoritative directives in the subjects' actions. There are problems with the no unanticipated information requirement as well, though I will not specifically address them here. If the problems with establishing the presence of autonomous benefits in the authority case are serious, this will be enough to show that the analogy does not hold: there will be no structural similarity involved. Before going into discussing these problems in detail, let me dismiss some preliminary objections first. This is important: it will show that drawing the analogy is *not obviously implausible*.

2. 2. Preliminary Objections

2.2.1. Two levels of decision

A preliminary objection draws on the fact that what makes impossible to form a rational intention to drink the toxin in TP, is the belief that the action, resulting from executing that intention, will be irrational. One cannot form a rational intention to do something he believes at the time of forming that intention to be irrational (this is an application of

[256] More formally: there is "a constraint on rational, deliberation-based intention. If, on the basis of deliberation, an agent rationally settles at T1 on an intention to A at T2 if (given that) C, and if she expects that under C at T2 she will have rational control on whether or not she A, then she will not suppose at T1 that if C at T2, then at T2 she should rationally abandon her intention in favor of an intention to perform an alternative to A." Bratman (1998: 62).

Bratman's linking principle). The objection is: no such belief is present in the case of legitimate authority. If one believes that authority is justified, one will believe that acting on its authoritative commands is rational as well. So, deciding (which involves intending) to follow authority is not irrational.

In reply: it is true that in the TP case the belief that one's particular action will *certainly* be irrational makes forming rational intention for that particular action strictly impossible. Nevertheless, there is a sense, in which *a belief* that certain particular actions will inevitably *be irrational*, even if commanded by a generally legitimate authority, is also present at the time of deciding (and forming a rational intention) to adopt the rule to follow authority. The question is why, nevertheless, it is held not to be irrational to decide (and form an intention) to follow authority in general, while the intention to drink the toxin cannot be formed rationally? If it will not be irrational to decide in this way, the TP would not apply to cases of deciding to follow authority.

There is a caveat here. One may resist the analogy because he believes TP applies only to cases where it is impossible to form rational intentions for *particular actions*. The decision to adopt a rule to follow an authority is not such a case: the decision is rather to adopt a *general* policy to act in certain ways. Since decisions to adopt a policy or a rule, or to follow authority in general, do not involve forming intentions for singular acts (the presumed by this objection domain of the application of the puzzle) these latter cases fall out of the reach of the TP.

This may suggest we have to relocate the application of the TP to the case of deciding to follow authority in particular cases. Thus the problem is not primarily to explain *how it is rational to decide to adopt a rule* to follow authority. The analogy may, rather, be at the level of establishing *whether it is rational to form an intention to follow the authority in a particular case*, when following the authority in a particular case will not causally determine the success of the overall policy of following authoritative directives generally.

In sum: there are two levels of forming a decision. The suggestion is that they need carefully be distinguished, because the TP may show up at one of the levels without necessarily appearing at the other.

Thus, let us distinguish

1. Deciding to adopt a policy to follow authority in general. The justification for that decision is *instrumental*: one is justified to take such a decision to adopt it *iff* one does better overall in terms of his ex ante reasons if one adopts it than otherwise.
2. Deciding (and forming an intention) to follow authority on a particular occasion. The justification here is content-independent. It is based on the *fact* that one has adopted a rule to follow authority in general, rather than on the instrumental justification for adopting the rule itself.

2.2.2. TP in Rationally Intending Plurality of Acts?

Consider the first, general, level of deciding to adopt a policy. The objection is that at this level the puzzle does not appear. Even though an instrumental justification for deciding is present, the intention to be formed is not an intention for singular act, and forming an intention for a singular act is thought necessary for having a TP type case.

Making a decision implies forming an intention. Thus the *decision* to adopt a policy will imply forming an *intention* to adopt a policy. A*dopting* a policy, however, is itself identical with *intending to follow through* with that policy. So, the intention (involved by the decision) to adopt a policy seems redundant ("intending to adopt a policy" would be the same as "intending to intend to follow through with the policy" directly, i.e. without the intervention of further deliberation).[257] The decision to adopt a policy is, then, simply adopting it. Adopting a policy implies intending to follow through with it. This adopting of a policy will, however, by the nature of the policy as applying to a plurality of cases, involve forming an intention to follow that policy on all cases of no unanticipated information. This, of course, presupposes that one can speak of intending plurality of acts, and not only of intending singular acts. I see no reason why intention in general should be restricted to intending singular acts only. Support for my argument here is

[257]An explanation why it seems strange to intend to intend may be that intentions are practical, i.e. necessarily directed at agency. One can only intend acts, courses of action, etc. ,Pink (1996: 18), and adopting a plan could hardly qualify to be an act. To adopt a plan is to form an intention to follow through with the plan. Certainly it is odd to allow for having an intention for forming an intention, as it is odd to allow for taking decisions for taking decisions. This does not imply, however, that one's more general intentions do not allow for having a hierarchical structure of subservient intentions – my intending an end may involve (though probably not "imply" *stricto sensu*) having or forming an intention for (what I believe to be) the necessary means to that end. It is not here the place to enter the intricacies of the debate whether intending an end implies intending the necessary means to that end. What is important for my argument is only that "intention to form an intention" is a redundancy.

provided by Bratman's work on intentions, where they are presented as allowing one to coordinate one's own actions over time (as well as to coordinate inter-personal actions) through adopting different plans, which suitably mesh together. This role of intentions presupposes that one can intend plurality of acts, since the execution of plans involves plurality of temporally extended acts.[258] Thus intending plurality of acts is not outlandish. Furthermore, unless TP is restricted for some special reason to cases of forming rational intentions for singular acts, it may in principle apply to cases of adopting policies, even though the implied by that adopting of a policy intention is an intention for a plurality of acts. I deal with this complex issue in the next section.

2.2.3. ER Solution to TP on Particular Occasion

At the second level, the intention to follow through enters the deliberative process,[259] in which a decision to follow *on particular occasion* is made. Only here do we encounter the autonomous reasons mentioned in the beginning of the chapter. The decision to adopt a general policy, taken at the first level, provides at this second level a protected autonomous reason for following through with this policy on a particular occasion. The second decision is based both on the exclusionary second-order and on the new first-order reasons provided by the preceding decision (or, here, authority), jointly comprising the protected reason. In cases of no new information, these two will be conclusive within the jurisdiction of authority: they will require acting as authority commands (or acting as initially decided: the decision is to follow authority).

The specificity of the autonomous reasons, according to Raz, is that they do not transmit the justification for their creation to the required action. This is the break of the transitivity of justification, involved in CiR, discussed in the first part of this thesis. Autonomous reasons (due t their CiR component) are opaque: the underlying the decision to follow the rule/authority reasons, are *not reasons for* following the rule/authority. Rather, the reason to follow the rule/authority is that the rule/authority so requires. This break of transitivity of justification implies that *the instrumental justification for deciding* to follow authority will *not* be transmitted *to justifying acting* as the authority requires.

---

[258] Bratman (1987).

[259] "…[intentions] are conduct-controlling pro-attitudes, they have inertia, and they serve as inputs in further practical reasoning." (Bratman, 1987: 27)

Consequently, here the TP cannot presumably arise, since *no instrumental justification* for intending the particular action is *directly* present (recall that presence of instrumental justification is a structural feature of this puzzle). *The presence of the autonomous reasons screens out the possibility of the puzzle appearing at the second level of deciding to follow authority on a particular occasion.*

2.2.4. Reappearance of the Puzzle?

There are problems, attending this autonomous reasons solution: they may well lead to the reappearance of the puzzle.

The main problem is to explain what accounts for the normative force of these autonomous reasons: what allows them to make practical difference to the ex-ante reasons, while still rendering acting on them rationally justified? For Raz,[260] recall, the normative force of a reason is a function of the evaluative characteristics of the action, which that reason recommends. In the case of autonomous reasons,[261] their normative force cannot rely on the value of the action they recommend, since they by definition are content-independent, i.e. do not depend on the content/the merit of the prospective action. On the other hand, their normative force does not depend on the justification for adopting the particular rule/decision to follow the particular authority either. The justification for the rule (which rule gives one a content-independent autonomous reason) is itself content-independent: does not rely on the value of the rule itself, but on the value of having rules, since it might be better to have some rules than not to have any.

For Raz,[262] normative force (what should be done) *always* ultimately relies on the evaluative (what is good about doing it). So, the ultimate justification for why one should follow through with the rule/authoritative requests will be that thereby one does better overall: it is instrumental. The idea is that though the autonomous reason does not show the desirability of the recommended action on its face (it is opaque), *the desirability of acting* as required by the autonomous reason still lurks somewhere in the background, and will be of instrumental character. Notice that what lurks in the background need *not*

[260] Raz's position of defining the normative (what one ought to do) in terms of the evaluative was considered in somewhat more detail in part one of my thesis.
[261] For this specific problem, see Raz (2001a).
[262] Raz (2001a: 15)

be the desirability of the action itself: the action itself might be entirely undesirable. Still, it might be *desirable to perform that action*, because one does better overall by acting as the rule requires (by acting on the autonomous reasons), than one would by acting directly on the reason, provided by the value of the prospective action.

The question is: is one allowed *this comparative judgement* (whether one does better overall by always following authority, or, rather, does better by generally acting as authority commands and acting only on this particular occasion independently of authority) to affect the relative desirability/undesirability of performing the act required *at the stage of deciding whether to follow authority on that particular occasion*.

Raz's solution is to introduce at this stage ERs as necessary components of the autonomous reasons (allowing them to screen out the possibility of the TP appearing here). The exclusionary reasons do this by *making the appeal to the justification* (the underlying reasons) irrelevant, since they anyway preempt *acting* on this comparative judgement. This "pre-empting" move is warranted, according to Raz, since going back and evaluating the underlying reasons as grounds for action, would involve their objectionable "double counting": these are the reasons, after all, on the basis of which the authority was meant to base its own judgement[263] and adjudicate between them. Reintroducing them at the stage of deciding whether to follow the authoritative determination on a particular occasion would mean that those reasons are allowed to count twice. Furthermore, Raz argues, disregarding in deciding how one should act[264] such anyway objectionable comparative judgement, is precisely *the only way* a better conformity to reasons could in principle be achieved.

The worry of those[265] who doubt the coherence of Raz's exclusionary reasons solution, however, is that unless one is allowed to appeal to the relative desirability of sticking to the rule/authority or acting on the underlying reasons, in deciding whether to act on autonomous reasons, one may end up acting irrationally. The mistake of the

[263] As it is required by the Dependence thesis, Raz (1986: 42-53), according to which authoritative directives should be based on the balance of the underlying, or ex ante reasons, which independently apply to the subjects of the directives.

[264] But not in one's reasoning: recall that ER is reason for action, not belief, and accordingly, it only precludes acting on the underlying reasons, and does not preclude considering the desirability of such action by balancing the underlying reasons. I raise some, admittedly inconclusive concerns against this distinction in Part one of this thesis.

[265] For such worries, see Moore (1989), Hurd (1999: 85), Alexander (1990: 12), among others.

rule/authority might be so great that it might turn out at the end to upset any benefits to be gained even in the long run from sticking with the rule/authority. The opponents' claim, thus, is that it is not rational to treat any reason as protected (because it is not rational to exclude considering the relative advantages of sticking/failing to stick to the rule/authority in deciding how to act on particular occasions). Following authoritative directives on an occasion is only rational, accordingly, if one need not treat them as protected and can bring in the instrumental considerations, screened out with the use of ER machinery.

Raz's response is that in abandoning the protected reasons account of authoritative directives, one could at best provide an account of what conformity to authoritative reasons means, but not what compliance with them is. His reply, then, is that his critics could at best explain "following authority" in the former, weak sense, and not in the latter, genuine sense.[266]

Be this as it may, the conclusion at this stage allows us to go back at the level of deciding to adopt a policy of following authoritative directives and try to build a case for its analogy to the TP case. So, if Raz's opponents are right and it is indeed not rational to treat on any particular occasion the authoritative directives as protected reasons with exclusionary force, one might not be able to rationally adopt a rule/rationally decide to follow authority, if following a rule/authority implies knowing that one is to be guided by protected (having exclusionary force) autonomous reasons.

Restatement of the "Toxin Puzzle" of deciding to follow authority:

We might be blocked from getting the benefit of improved conformity to reason by the impossibility to form a rational intention to treat the autonomous reasons provided by the authority as protected (having exclusionary force).

[266] Interestingly, this is readily admitted by Regan (1989: 1095), for example. In arguing against the rationality of the strategy of always deferring to authority, he concludes that the rational thing to do might indeed be to "follow authority." But to follow it not in the sense of deferring to it (treating it as practical authority), but in a sense of using it as a source of information (treating it as a source of "indicator rules") about the best course of action, which may or may not coincide with what authority actually commands. One's most rationally justified indicator rule may be – follow authority when it is sunny, but never on a rainy day! Rationality, for Regan, may warrant following authority in a weak sense only - not as practical, but as theoretical alone. One wonders whether what Regan describes as an indicator rule account of authority could be an account of authority – one seems to be required to follow reason only.

With this last move, I have finally set out the aspects of the case of rationally deciding to follow authority, which most closely analogize the TP case. Of course, the plausibility of the analogy is conditional on the warrantedness of the conclusion, that treating authoritative directives as autonomous protected reasons with exclusionary force is irrational, and this claim has not been established. Nevertheless, what is important here is that this allows us to go as close as one could to drawing the analogy of the TP case to that of deciding to follow authority. What remains to be seen, then, is whether there is indeed a structural similarity between these two cases.

3. "Autonomous Benefits" in Deciding to Follow Authority?

A structural feature of the TP is that it involves what Kavka calls autonomous benefits: the benefit of forming the intention to drink the toxin (and thus getting a reward) is autonomous with respect to (not causally dependent on) actually executing that intention (i.e. drinking the toxin). The action (the execution of the intention) is temporally located in such a way, that it *cannot causally determine* getting/not getting the reward. Since that fact (that the benefit, the only justification for forming the intention, is causally independent from actually executing the intention, and does nothing to justify performing the action itself) is known to the agent, he is prevented from rationally forming an intention for the act so characterised. The requirements of instrumental rationality prevent the agent from getting the reward. The autonomous benefit feature is the culprit.

As stated above, the rationality of deciding to follow authority is justified by instrumental considerations: doing better (better conforming to ex ante reasons) in the long run. But the benefit of doing better in the long run, it will be objected, is not autonomous with respect to actually following the authority. Following authority causally determines that one does better in the long run. This consideration in my view is the *strongest argument against the analog*y.

It might be possible to rebut this argument by maintaining that the justification (doing better in the long run by following authority) for taking the decision to generally follow authority *does not causally depend* on following the authoritative directives *on each*

occasion. My interest is again to see whether the analogy could be drawn at this general level of deciding to follow authority.

The problem, already alluded to above, for this way of establishing the analogy (by appealing to the fact that receiving the benefit of authority does not causally depend on acting as required on each occasion) is that TP presumably arises in one-shot cases only. Only if a benefit would be received on the basis of forming an intention for action, and receiving it does not depend causally on the subsequent to getting the benefit action, is the benefit strictly autonomous. The explanation for the impossibility to form rationally the required intention in the TP case, is that since the action is temporally located *after* getting the benefit, this action *in principle cannot causally control* whether the benefit is given. This strategy does not do justice to a "basic fact about our agency…along with a change in temporal location normally goes a change in the agent's causal powers. What is up to the agent is what to do from now on." Bratman (1998: 66)[267]

The alleged autonomous benefit of doing better in the long run in the authority case, however, will not be *in principle* causally independent from acting as the authority demands. The benefit is *extended over time*, and is temporally "located" in a way that the action (singularly or cumulatively) *could* in principle causally determine it.

For the analogy to hold, we need a case where getting the benefit of doing better in conforming to reasons is *in principle* causally independent from the action of complying to authoritative commands. This could only be true, if at all, in a one-shot case. Further advantage for having a one-shot case, is that for the analogy with the TP case to hold, a *no unanticipated information case* is required. This is most easily satisfied in one-shot cases, where uncertainty for the future need not enter the picture.

So, could the temporally extended case of the instrumental justification for deciding to follow authority be reduced to a one-shot case, so that an analogy with the TP case could plausibly be drawn?

3.1. The Backward-Induction Argument

[267] This is Bratman's explanation why McClennen's resolute strategy for dissolving TP does not succeed. The resolute strategy suggests that in deciding whether to follow through with a plan, one evaluates courses of action, not singular acts. Part of these courses of action may already be in the past, and thus not be in the causal control of the agent, but may still, nevertheless, determine how the agent should act.

A way to reduce the extended over time case of following authority into a one-shot case, is to apply to it backward (recursive) induction. The backward induction argument [268] was introduced in discussions of the rationality of reciprocation. It may as well help in the analysis of the rationality of deciding to follow authority, by helping bring it close to TP.

Backward induction in deciding to follow authority:

1. Assume full knowledge of one's own rationality (a standard assumption in backward induction arguments)

2. Assume the truth of Raz's claim: The benefit of improved conformity to reason depends on the practical difference authority makes as a mediator between the persons and their ex ante reasons. Authority can make such practical difference, only if its directives are taken as providing protected reasons for action with exclusionary force. If authority is treated as theoretical only, as providing some reasons for belief to be balanced by the subjects themselves against the ex ante reasons they have independently of authority, it cannot bring such improved conformity to ex ante reasons

3. Assume (unrealistically) there is a known last occasion (one is, after all, a temporally finite being), when one considers whether to follow authority's command, even though it conflicts with one's all-things-considered judgement concerning the merits of the required action. On this last occasion, then, one deliberates whether to take the authoritative directive as a protected reason for action with an exclusionary force.

4. The benefit of *generally* following the authority and thus achieving better conformity to reasons in the long run cannot causally depend on following authority on this last occasion (since it is the last – no future short or long term benefits are expected). The benefit is autonomous with respect to following authority on this last occasion.

5. It is irrational to act directly on the protected reasons provided by the authority on that particular last occasion (the general benefit is autonomous with respect to this particular act) without considering the relative benefits of sticking to authority or balancing the ex ante reasons alone instead (from 2. and 4.). The benefit of sticking to authority on the last

[268] This argument was presented by Duncan Luce and Howard Raiffa (1957:97-102), and was further developed in "The chain store paradox" by Selten (1978: 127 –159). The references are from Sugden (1992: 201).

occasion would depend only on one's mistrust of one's own judgement on the balance of reasons in this particular case. In case of strong disagreement with authority, given one knows there are no perfect authorities (great mistake is always possible), the rational thing to do is to decide on the balance of *all reasons*: one's own judgement, the reasons to trust the authority on this particular case, etc. One treats authority on this last occasion as theoretical only, and not as providing protected reasons with exclusionary force.

6. Since it is irrational to act on the protected reason directly on the last occasion, it will not be rational to act on it on the-first-before-the-last occasion. This is so, because there are *no future benefits* that could be gained by following authority, which could be *causally determined by that act*. This follows from 1. - full knowledge of one's rationality: the-first-before-the-last self knows that his subsequent self is rational, and will not follow authority as a practical authority[269] on a last occasion - no further benefits expected. No expected benefit for the first-before-the-last self as well, then: if he is rational, he would not follow authority simply because it so requires. The same applies to the second-before-the-last self, etc…

7. When the zipping back is complete, we reach the one-shot case of deciding whether to follow the authority. Since the last in the backward order/first in the temporal order self knows that all the decisions of the subsequent selves to follow authority will be rational and there will be no single act of following authority as an authority, the benefit of conforming better to reason by following it cannot lie in the future for the temporally first self as well. He stands to gain nothing by deciding to follow authority instead of his reasons directly: it is not rational of him to decide to follow authority, simply because authority so commands. If it would ever be rational in this situation to decide to follow authority, it would be so only if the benefit of better tracking reason does precede and thus is autonomous with respect to subsequently acting on that decision. This is so, since the justification in my artificially constructed example for deciding to follow authority cannot come from subsequent to the required by the authority complying action: there will be no more [and this is known by the deliberating agent] complying actions, and thus no further benefits flowing from them. The puzzle arises because it will not be possible to

[269] The agent here might follow authority, but not *because* it is an authority. Rather he might follow it because the balance of reasons recommends so.

rationally form the implied by that decision to follow authority intention, since the benefit (better conformity to reason), if present, will be known to be autonomous (causally independent) with respect to actually acting as intended.

In this way, the "TP" may show up in the case of authority, claiming to provide protected reasons for action with exclusionary force.

3.2. An Objection Considered

Before evaluating this claim, let me address an interesting objection against using the backward-induction argument for the case of deciding to follow authority.[270] It runs roughly as follows. Though it is true that one's decisions can be interdependent (either because one's resoluteness in the future may be affected by one's present resoluteness, or because one's decisions to act carry out consecutive steps in an on-going plan), it might still be rationally unobjectionable to refuse to comply with authoritative directives, and this may happen in any (not the last!) of the particular cases of deciding to follow authority. This is so, because authoritative reasons are conclusive (protected), but scope-limited and there might be clear criteria for limiting their scope. If so, it will be rational to follow a rule with the structure: don't follow authority if X, and follow it in all other occasions (non X). If one is strictly guided by this rule, not following authority on a particular occasion (X) will not affect one's resoluteness to follow it later (non X). Admittedly, it might be difficult to draw a clear distinction between X and non X, but even if so, this points to a different problem than the one I am addressing. Problems with drawing such a clear distinction point to the *instability of a strategy* to decide to follow authority, and are *not related to the puzzle about instrumental rationality* I discuss here.

I agree with this last point. But it is not entirely clear to me whether this objection applies to my argument. Firstly, and most importantly, let me stress that there is a distinction between the case of following authority as it directs one to follow it (it claims to provide protected reasons for action, where it is authority itself and not its subject that claims the right to determine in each particular occasion of disagreement where the limit of its directives lies) and following a rule, where one himself has conclusively determined the

[270] This objection was raised to me by Professor Kis.

scope of the rule, and the rule itself does not "claim" to be determining it on each occasion of disagreement.

This clearly points to a *dis-analogy* between the *case of deciding to follow any rule*, and that of *deciding to follow authority* in particular.[271] In the latter case authority's claim always to preempt one's judgement and direct one's action conclusively, may present us with a puzzle about the rationality of deciding to follow it, which is not reducible to the instability problem, common to both. By speaking of "adopting a *rule* to follow *authority*" and concentrating on specifying the conditions for the applicability of the rule itself, one may conflate "adopting a rule with [externally, i.e. in advance clearly and conclusively settled] limited scope of application" with "deciding to follow authority, [which always claims for itself the right to decide where the limits of its orders lay and does not allow external to it delimitation of its scope]." This conflation accounts for the initial plausibility of this objection, but removes the sting from my argument at the expense of not providing an adequate description of what is crucially at stake in the case of deciding to follow authority (at least on Raz's account of authority). This conflation may be due to the preeminence of the point of view of the subjects in the above examples, at the expense of neglecting the authority's point of view. [272] So, even though it is true that it is the subjects that decide whether to enter the authority relationship, once in it, it is the authority (law) that determines the actual terms through its claim to normative supremacy over all other normative domains: it claims to have unlimited authority,[273] a claim not made by mandatory rules.

Secondly, the above objection is supposed to work if it managed to establish that decisions to follow authority on each occasion are not inter-dependent – what I now

[271] This is a point that needs further development. Though reasons provided by authority and reasons provided by rules share some structural features: both are protected reasons with exclusionary force, authority, and law in particular, further adds to this common structure a *claim to obedience*. Rules clearly cannot make such claims. I already touched upon these issues when discussing problems with authority's/law's comprehensive claim to supremacy.

[272] For a suggestion along these lines, see Regan (1989: 1018). The advanced by him indicator rules conception of authority takes the point of view of subjects as having normative primacy, and he somewhat inconclusively suggests that the difference between his account and that of Raz's stronger understanding of "authoritativeness" might be due to the latter author's preoccupation with the point of view of authority.

[273] "Law provides ways of changing the law and of adopting any law whatsoever, and it always claims authority for itself. That is, it claims unlimited authority, it claims that there is an obligation to obey it whatever its content may be." Raz (1986: 77) More to the point, see Raz (1989: 1069): "While all legal systems allow for certain moral defenses and exceptions, they claim the right to determine which moral defenses and exceptions count…"

decide does not affect how I will decide later. Hence, there is no strategic game between one's consecutive selves, and between those and authority, which presumably allows for running the backward induction. Notice that the picture this objection draws of one's relation with authority is static: unconnected and not temporally extended, one-shot interactions. This is in contradiction with the picture of authority relationship I discuss, which is aggregative, not limited to one-shot cases of interaction, and extended in time.[274] This is indeed Raz's own view:

> "Usually arguments for authority are general. They apply to the justification of the use of public power over a range of issues, for an *extended period of time*." Raz (1986:73, emphasis added)

A note of clarification is due here. It will be objected that aggregation need not logically entail temporal extendedness of multiple cases, but could rather be understood in probabilistic terms: the authority might stand a 99% percent chance of being correct. I agree with this. However, let me point out, that running a-temporal, counterfactual comparisons (apart from problems with determinacy of the results) is not the way humans gain knowledge about the relative superiority of authority in terms of expertise, capacity to solve co-ordination problems etc. Note that, lacking such knowledge,[275] one cannot rationally submit to authority. The normal way humans gain such knowledge is either by relying on their own experience in sufficiently long, extended in time interactions with authority, or on the testimony of others. Such is the case of the exceptionally healthy person, who only once in his life needs medical help, and goes to the doctor, presumably because he trusts that he is a good doctor, which trust presumably is warranted again because of the favourable temporally extended experience of others with this particular authority. My conclusion is that in the world as we know it, the rationality of submitting to authority depends on and presupposes a temporally extended interaction with authority.

[274] Recall the discussion of the different interpretations of NJT in Chapter 2: I distinguished cumulative and one-shot interpretations and claimed that the cumulative seems better to support NJT, especially given that legitimacy should allow authority occasional mistakes.

[275] I do not mean certainty, acquired through careful investigation of all the merits/demerits of following the authority in question. I guess some weaker, much less reflective cognitive state would suffice, although I am not capable of specifying in precise terms this minimally necessary condition.

Furthermore, and relatedly, my interest is in the compatibility of the instrumental justification for deciding to follow authority with authority's nature, and this justification, being cumulative, again presupposes temporally extended, multiple interactions with authority. By failing to comply with authority, which claims that condition non-X is met (and by instead following one's own rule to only comply with authority if one himself believes it is non-X, thus disagreeing with authority on this particular occasion), one demonstrates that one is not in an authority relationship anymore – but then the cumulative benefits, of which the instrumental normal justification thesis speak, would not be accessible either. Note that it is cumulativity of the benefits, coupled with one's uncertainty as to whether one's actual following authority on an occasion would add to or take from the positive balance, which allows for the possibility of having autonomous benefits in the case of authority (on this feature I will spend more time later in this chapter). Besides, Raz himself explicitly uses an argument from "inter-dependence" to prove his case for the duty "to stop at the red light in the deserted intersection." If one is to ascertain whether all the conditions for not stopping are met, one would need to do it on many other occasions, thus foregoing the instrumental benefit of authoritative guidance.[276]

Going back to the Toxin Puzzle. To avoid it, it seems, one would need to maintain that the benefit of following authority is essentially future-oriented and, because of this, cannot *in principle* be autonomous with respect to the complying with the authority particular action. However, if the backward-induction argument is sound, it establishes that one cannot rationally expect any future benefits from following authority simply because authority so commands: the benefits, if any, would seem to be autonomous. However, since the presence of autonomous benefits is what generates the puzzle in the toxin type (and poses problems in reciprocation cases, the primary application of the backward-induction arguments), one will in the discussed above case as well be rationally prevented from getting the instrumental benefits. A rational deliberator could

[276]Raz (1979: 24 –25). I will discuss again this case when the problems around the distinction clear/great mistake are addressed.

not get any benefits, which authority could bring, if treated as a practical authority. Or so it seems.

3.3. Limits of Knowledge and Rationality: Non-Autonomous Benefits

It will be immediately noticed, that the argument above is not applicable to the circumstances, in which authority will indeed be instrumentally beneficial. The argument only works if full knowledge of one's instrumental rationality is assumed. These are conditions, when following authority may indeed be irrational. The benefits of following authority (which could either directly bring improved conformity to ex ante reasons, as required by NJT, or could be beneficial in facilitating the process through which decisions are reached[277]) are available precisely under conditions of limited knowledge, temporal limits on individual reasoning capacity, limits of our resolve (weakness-of-will cases) as well as limits of our collective rationality, imposed by our maximizing individual rationality (PDs and coordination problems). Only the last benefits (overcoming weakness of will and securing solutions to PDs) may still be available in conditions of full knowledge of rationality and thus may justify following authority.

The claim, then, is that once the limits of knowledge, reasoning capacity, time resources, the psychological costs, etc. are taken into account, the benefit of following authority need not be construed as always necessarily future-oriented. The generally improved conformity to reason is an overall, usually long-term benefit, which could be received by following the authoritative directives. It should be noted that one stands to gain even on the spot, from following authority by taking its directives as protected reasons with exclusionary force, even (1) when one disagrees with authority as to whether the required action is good on balance, and even if (2) one turns out ex post facto to have been right and the authority wrong. One's benefit here will be in terms of economizing on efforts to gain knowledge, efforts to apply one's reasoning capacity, when time is limited and decision is due, psychological costs, etc. These benefits may outweigh the cost of authority occasionally commanding wrong actions (but not if the authority systematically commands wrong actions). Such economizing benefits need not be connected with

[277]NJT, I claimed in chapter 2, is intended by Raz as an exclusively substantive test of legitimacy, which does not exclude, however, favouring procedures when they are instrumental to bringing best outcomes.

expecting further future benefits from sticking with that policy of economising on decision-making costs.

One should not be misled to conclude, however, that denying that some benefits of following authority are future-oriented amounts to concluding that they are *in principle* autonomous. The local benefit of economising here, *though not future-oriented, is not autonomous either*: it *causally* depends on actually treating the authority as providing protected reasons for action and on actually acting on them. Further, receiving this local economising benefit may depend on trusting that authority is being overall beneficial by getting the balance of ex ante reasons more correctly than would the individual subject, if left alone.

4. Autonomous Benefits in Deciding to Follow Authority

4.1. The Wrong Belief Case

The benefit of deciding to follow authority (reducing decision-making costs in particular) may be *autonomous with respect to actually following* the authority in cases when the subject *mistakenly believes to be complying with the authority, while in fact failing to do so.*[278] Thus one will not waste decision-making resources on balancing the reasons for and against the commanded action, and will believe to be acting as commanded, without necessarily doing so. The subject, paradoxically, could on occasion even get a double benefit as well: if he mistakenly believes to be complying with an authoritative directive, which turns out to be misguided, but acts in the right way by not actually complying in one's actions with this directive. He will get the "economising" benefit of the decision to comply with authority, rather than act on one's own judgement of the merits of the case, together with the benefit of ending up acting on the right balance of reasons.

Obviously, this is an atypical case, not central to Raz's account of authority - it is not covered by NJT, and would be a too slim basis for establishing the analogy with the TP case. Furthermore, though there is here presence of an autonomous benefit, this case is not analogous to the TP case, since the belief that one will not have a reason to act in

[278] See Raz (1989: 1161). This is the case of the lucky father, who promises his wife to disregard his own interests in choosing the best college for his son. Believing to act on his promise, he may actually choose the second-best college (believing it is the best) which as a matter of fact will serve his interests as well (it is cheaper, and he needs the money to retire and write a book). The promise for him was indeed an exclusionary reason, even though his action did not actually comply with that reason.

executing the beneficial intention is not present at the time the decision is taken: for all the agent knows, he will have such a reason, so there is no puzzle involved here. Even if it were strictly analogous, it would not be sufficient to establish the analogy I am interested in: it requires showing that deciding to follow authority generally brings benefits, which are *in principle* autonomous with respect to following its directives and not contingently so (through a mistake!), as in the above case.

4.2. The Ambiguous "Clear Mistake" Case

There might be, however, more interesting cases, when the benefit of following the authority could be autonomous with respect to acting on the authoritative directive, not on account of the wrong belief of the subject that he is complying with it, but even if the subject knowingly abstains from complying with it. This could be the case when authority makes clear mistakes. The benefits of economizing on knowledge and reasoning could still be gained even if one does not follow the authoritative command in acting on those occasions, when authority requires *clearly* wrong action. Since the mistake is clear, it presumably "does not require going through the underlying reasoning process,"[279] and so does not require any special effort (gaining knowledge or burdensome application of reasoning capacity) to be discovered. It would seam, then, that clear mistakes on the part of authority defeat its claim to be obeyed. If that was so, we could have a second case of autonomous benefits. And, note, that we could here draw a plausible analogy with the TP case. Since in deciding to adopt a policy of always following authoritative directives one knows, that when those directives are clearly wrong, and there are no future benefits to be expected,[280] one need not follow them, one might be prevented from rationally deciding to adopt such a policy.

This raises the complex issue whether for Raz it is acceptable to disregard authoritative directives when they are clearly wrong. Though the fact that he draws a distinction between clear and great mistakes points in that direction, he avoids taking a position on

[279] Raz (1986: 62)

[280] Of course, for the analogy to hold, it should be presumed that the agent knows, in not acting on clear mistakes, not only that this is a case of clear mistakes, but that there will be no future benefits as well, and he should be aware that he will have this knowledge already at the time of deciding to adopt the policy. This is a strong assumption: while one will have a knowledge about the clear mistake, it is very unlikely that one will be certain that no future benefits of following authority lay ahead.

that.[281] Besides, the already mentioned discussion of "the red light at the deserted intersection" case shows that Raz is not prepared to exclude clear mistakes from the scope of application of authoritative directives, and for good reason. Though there is a strong pressure towards excluding them from the requirement to comply with authoritative directives, in order to avoid charges of irrationality, this would come at the expense of providing an unacceptably "intermittent" picture of authority, not true to our (both everyday and philosophically enlightened) understanding of authority as requiring "blind obedience."[282]

Relatedly, accepting that clear mistakes put such directives outside the jurisdiction of authority would contradict Raz's Preemption thesis. It says that the authoritative directives displace the ex ante reasons: it excludes acting on those reasons, even when subject's own judgement about the balance of those reasons is right and authority's wrong. According to this thesis, it does not matter whether authority's mistake is great or clear. This is, anyway, irrelevant, because the authoritative directive is content-independent: one has to act on it irrespective of any substantive merit or lack of such.

The issue of clear mistakes helps illuminate, I believe, the problem of having fixed jurisdictional boundaries of authority (important for the plausibility of the Preemption thesis). It is also relevant for the "instability problem" of the decision-making strategy. In the next part I will need to come back to these issues.

While the case of authority making clear mistakes could in principle warrant drawing the analogy with the TP case - were Raz to admit that clearly wrong directives fall outside of

[281] "Even if legitimate authority is limited by the condition that its directives are not binding if clearly wrong, and I wish to express no opinion on whether it is so limited…" Raz (1986:62). Regan (1989) finds a fault with this Razian ambiguity. Shapiro (2002a: 405), on the contrary, takes it that Raz excludes the case of clear mistakes from the obligation to comply with such directives. I think they both misrepresent Raz's position. Raz is clear on that "The wavering that [Regan] chides me for in this regard [should an account of the concept of authority be committed to agent-neutral consequentialism], as well as that concerning the question *whether a clear mistake puts a directive outside of jurisdiction of the authority*, is no wavering at all. I was putting forward an account, which explains a concept used by people holding different views on these issues. To make it a good account, I had to recognise that, and avoid any explanation that takes sides on these issues". Raz (1989: 1184, emphasis added)

[282] This objection against the instrumental accounts of authority as leading to such "intermittent," patchy pictures of obligation, is raised by Dan-Cohen (1994: 33-34). It was partly discussed in Chapter 3, where the disjunctive view of normativity and coercion was analysed.

the jurisdiction of authority, which he does not,[283] this is not so for what is considered the much more interesting and troubling case of authority making great mistakes. On this Raz is explicit.[284] It has been suggested that the bindingness of greatly mistaken directives is what raises serious doubts concerning the coherence of the Razian account of authority, since instrumental considerations [the normal way authority is justified] do not warrant acting on such greatly mistaken directives. Irrespective of the success of this charge, it will not serve to establish the sought-for analogy with the TP case. The reason was already demonstrated: the benefit of improved conformity to ex ante reasons is not autonomous, or causally independent, with respect to actually acting as commanded - both when the commands are mistaken and when they are warranted. If there is a problem with this position, it is not in the alleged analogy with the Toxin Puzzle case.

5. Conclusion

My conclusion is that the analogy of deciding to follow authority with the Toxin Puzzle case does not hold. The limits of knowledge, rationality, decision-making capacity, resolve, etc. stand on the way of drawing such an analogy. While it is true that there are cases, in which deciding to follow authority involves "autonomous benefits," not causally dependent on actually acting on this decision, since they are not known to the agent, they do not necessarily threaten his rationality. For the analogy to hold, we need more than just the presence of autonomous benefit. Knowledge on the part of the agent at the time of forming the intention, that at the time of executing the intention, (.e. at the time of the action), he will know that the benefit is autonomous, and thus will realize that he has no reason for acting as initially intended, is also required. Both the presence of autonomous benefits, and the awareness of them both at the time of deciding to follow authority and at the time of executing a concrete intention to follow authority as well, are missing from the case of treating authoritative directives as protected reasons with exclusionary force. A general awareness, that there will inevitably be cases, when acting as authority requires

[283] Those theorists, who advance instrumental accounts of authority, allowing for an intermittent picture of obligation, as Gans (1992), for example, should come closer than Raz in presenting the case of deciding to follow authority as analogous to TP.

[284] "…the [dependence] thesis is not that authoritative directives are binding only if they correctly reflect the reasons on which they depend. On the contrary, there is no point in having authorities unless their determinations are binding even if mistaken…" Raz (1986: 47)

is sub-optimal, is not sufficient. If there is a valid charge of incoherence against Raz's account of authority in terms of protected reasons with exclusionary force, it should be sought along different lines.

The discussion in this part, let me point out, was not entirely superfluous. The analysis of "autonomous" benefits will prove useful, I believe, for my discussion of the "instability" problem for an instrumentally justified decision strategy in the next chapter.

Chapter Seven

The Rationality of Deciding to Follow Authority: The Instability of the Instrumentally Justified Decision Strategy

In the preceding chapter I addressed the question whether the Toxin Puzzle case applies to the case of rationally deciding to follow authority on Raz's account. Though my conclusion was that drawing such an analogy is not warranted, the problems discussed with the possibility of having *autonomous benefits* may still present us with a problem about the rationality of deciding to follow authority. The problem is that if there is no stable rational strategy of following authority, it might indeed not be rational to decide to follow authority. Exploring these issues is the task of the present chapter.

It has the following structure. First the problem of instability of the rational strategy to follow authority is defined and its sources identified. After next pointing to an ambiguity, which makes unclear the success of a Razian strategy in that regard, I address the question whether Michael Bratman's strategy (allowing for rationally undertaking stable commitments) can be used for explaining how it can be rational to decide to follow authority, if following authority implies accepting its directives as protected reasons with exclusionary force. My conclusion is that there are serious problems with applying this strategy to the Razian model of authority. At the end of the chapter I bring in Scott Shapiro's alternative "Constraint" account of authority, which seems to hold the promise of providing the necessary stability in the decision-making strategy, as well as of solving the problems of the rationality of following authority. This is done by altogether abandoning the Decision model, shared by Raz and his critics. My contention is, however, that Shapiro's account comes dangerously close, and may ultimately collapse into one of the more problematic versions of the Decision model: the Resolute choice model. That this version is problematic is shown by the Toxin Puzzle – this puzzle was offered as a reductio ad absurdum precisely of this model, and construing one's relation with authority on this model would more closely analogize TP than Raz's own account of authority. Thus there is little to recommend this Constraint model. Abandoning the

instrumental framework for justifying authority seems to be the recommended route instead.

1. Defining the Instability Problem

1.1. Clear/Great Mistake Distinction – Ambiguity in Clear Mistake Cases?

A common charge against Raz's analysis of authority in terms of protected reasons with exclusionary force is that it renders following authority irrational, since it is allegedly irrational to decide to follow authority on a particular occasion without first checking whether the authority commits a great mistake.

In addressing this concern Raz distinguishes cases of great mistakes from those of clear mistakes. He claims that even if authority occasionally makes great mistakes, we might still be better off following its directives rather than figuring out for ourselves what to do. Recall, for him the Preemption thesis (the authoritative reasons replace, preempt our own reasons) naturally flows from NJT. Thus the decision-making costs might be such that they outweigh any advantage to be gained from avoiding the occasional great mistakes authority might lead us to make. Or the expertise of the authority might be such as to warrant following its directives, etc. The rationality of acting as the Preemption thesis demands, is the result of cost-benefit analysis: it is better to save on decision-making costs (time and resources spent on gaining knowledge and exercising our reasoning capacity) and thus act as the Preemption thesis requires, if the benefit of following the authority outweighs the disadvantage of occasionally committing great mistakes.

Raz's position is ambiguous, as pointed out in the previous chapter, concerning authority committing *clear* mistakes. They do not require any effort to be detected. Accordingly, conformity to reason could only be improved, without incurring extra costs, if one refrains from following authority in such cases. If this is not Raz's conclusion, then he should have important reasons for not putting clearly mistaken directives outside the jurisdiction of authority. It is difficult, however, to immediately spot them.

The "cost-benefit" analysis suggests that discovering a clear mistake does not require going to the underlying ex-ante reasons and doing all the work authority was meant to do before issuing its directives: clear mistakes simply "present" themselves without requiring any effort to be discovered and recognised as such. Furthermore, no

objectionable double-counting of the ex-ante reasons would be involved if such directives are dismissed: it is only ensured that the ex ante reasons are counted at least once, since authority is clearly wrong and obviously has not taken them into account.[285] Lastly, one need not be always alert to the possibility of mistakes, always go to the underlying reasons, etc., in order to discover clear mistakes: this allows for improving conformity to ex ante reasons without incurring extra costs in decision-making, knowledge, anxiety, etc.

One main reason why Raz nevertheless is reluctant to exclude such mistaken directives from requiring obedience, I believe, is in the perceived danger of infecting the strategy of always following authoritative directives with "instability." This is the reason he presents in establishing his case for obeying the "red light in the deserted intersection": if one is to ascertain whether all the conditions for not stopping are met, one would need to do it on many other occasions, thus foregoing the instrumental benefit of authoritative guidance.

However, an objection goes, if this indeed is a case of a clear and not a great mistake (as it is certainly the case here) then Raz's reasoning is not warranted. One need not *actively* "ascertain" whether *all the conditions* are met, since they here present themselves as being met. One only need "*passively" register*, and recognise that fact, without being involved in any thorough deliberative process. It is not obvious how such passive recognition would threaten getting the instrumental benefits of authoritative guidance. That thorough deliberation will be involved in discovering great mistakes is beside the point here.

This observation allows us to get to the main issue, raised by the case of clear mistakes: excluding them from the jurisdiction of authority would go against the Preemption thesis, and against the account of authority in terms of protected reasons with exclusionary force more generally. One main argument both for this thesis as well as for the account itself, is the *functional argument*. It maintains that unless authoritative reasons are always treated as protected exclusionary ones, they could not perform their function of ensuring improved conformity to ex ante reasons. That is, in order to perform their function,

[285] Shapiro (2002a: 414) makes a similar point in arguing against the success of the argument [for the Preemption thesis] from double counting. His suggestion, however, is stronger: making sure that not only clear, but great mistakes as well are avoided, thus defeating the Preemption thesis, may be warranted because of the justified concern that the ex ante reasons should be counted at least once.

authoritative reasons should exclude acting directly on the ex ante reasons, clear mistake cases included. Further, for the Preemption thesis to work, a relatively fixed jurisdiction of authority should be presumed: it should be possible to clearly settle in advance the issue whether a case falls within the jurisdiction of authority, before excluding acting on the ex ante reasons in the concrete case.

The presence of clear mistakes threatens this picture of clear jurisdictional boundaries. Suppose that in issuing directives within the class of cases included in its jurisdiction, authority happens to commit a clear mistake. If this amends the jurisdictional boundaries of the authority, this renders jurisdictional boundaries "soft." The logic of "exclusion" works only with more or less clear-cut jurisdictional boundaries. So, a clear mistake should not be taken to amend them. Rather, even the clear mistake of authority should warrant the exclusion of the ex ante reasons.

However, the functional argument, meant to justify the exclusion of ex ante reasons, is not sufficient to exclude clear mistakes. As already pointed out, conformity to ex ante reasons may only be improved if one does not act on the clearly mistaken directive, but follows his own obviously correct judgement on the balance of ex ante reasons instead. The functional argument for the Preemption thesis needs boosting.

The *argument from the "instability"* of an instrumentally justified, rational decision-making strategy to follow authoritative directives, while allowing for exceptions in cases of mistakes, could be such a boost. I take Raz's argument for obeying the law and stop at the red light in a deserted intersection, as relying on such type of argument "from instability." But does it support the Preemption thesis?

1.2. The "Instability" Problem

The problem of instability, I believe, is connected to the perceived possibility on part of the subjects of having "autonomous benefits" from following authority. Since one knows that getting the benefit of improved conformity to ex ante reasons need not causally depend on actually following the authority on each and every occasion, one *is tempted on each occasion of disagreement* with the authority, not to comply with its directives, but act on the balance of ex ante reasons as he perceives it. Note that this is indeed the rational thing to do in genuine cases of autonomous benefits. However, if this temptation

is always present in cases of disagreement, and one gives in to it often enough, the benefit of improved conformity to ex ante reasons will turn out *not to be autonomous* after all. Recall that this improved conformity benefit does indeed *causally depend* on taking the authoritative commands as protected reasons on *most if not all* cases of disagreement.

This may be taken to suggest that unless the protected reasons provided by authority are treated *as conclusive (absolute within the jurisdiction of authority) reasons* on each and all occasions, the point of deciding to follow authority will be defeated. This, recall, is the main argument for the Preemption thesis.

Treating protected reasons on all occasions as absolute (within their jurisdiction), however, is not rational: clear mistakes are a case in point. Further, there are the cases of exceedingly great mistakes, when following the authority may also not be rational: any advantage to be gained by following authority generally may be greatly outweighed by the disadvantage of following authority in the case of such exceedingly great mistake.

Then, the protected reasons may need to be taken as merely prima facie, i.e. as not absolutely protected by their exclusionary force. Authoritative reasons thus have great presumptive force, sufficient to decide the issues most of the time. They will, however, never fully exclude balancing the pros and cons of obeying. In this latter case the benefit of having authority arguably will again be lost.

This rationally sanctioned, and thus inescapable, fluctuation between treating authority as absolute (within its jurisdiction) and treating it as merely prima facie, renders unstable an instrumentally justified strategy to follow authority.

Put more formally,[286] the problem of instability is:

1. One either (a) decides always to follow Authority (treat A-reasons as absolute) or (b) decides to do so but to leave room for exceptions (treat A-reasons as prima facie).

2. It is not rational to treat A as absolute: (a) A may give suboptimal directives (b) The agent may see it at no great cost.

3. When 2. (a) or 2. (b) are met on an occasion, it is not rational to follow A.

4. But if it is rational to leave room for exceptions, then it may *never* be rational to follow the A (exclude one's ex ante reasons from consideration)

[286] I am grateful to Professor Kis for helping me see and present the problem in this crisp form.

5. The right decision strategy is to decide whether, all-things-considered, it is rational to do as the A tells one to do
6. This means that one fails to follow the A as A

The instability of the instrumentally justified decision strategy to follow authority is ultimately due to the fact that it is always a matter of *decision*, on the part of the *subject* on each occasion of disagreement with authority, whether the authoritative directive succeeds in preempting the ex ante reasons or balancing of those reasons is rather required by rationality. It is always a matter of an all-things-considered judgement whether to give priority to the reasons given by authority or act on one's own. But then one fails to follow authority as an authority, and follows one's own judgement instead. We are back with the rationality paradox of authority, pressed by the philosophical anarchists.

2. Is the Instrumentally Justified Decision-making Strategy Stable?

The way Raz can boost the Preemption thesis is by offerring an account of an instrumentally justified decision-making strategy that is „stable", where the explanation for its stablility is precisely that the conditions of the Preemption thesis are met there.
Raz does offer such an argument.[287] This is his argument about the maximally efficient way to treat the advise of an expert, better than its subject in reaching the right decision. Raz agrees that the subject will certainly improve his success rate, were he to take the expert's pronouncements as advise, to be added to the reasons, on the basis of which the subject will base his judgement how to act. In cases the advise tilts the balance and one decides as the expert would, one's success rate equals that of the expert. Not so when the advise is not enough to tilt the balance and one does not end up acting as the expert advises. However, even taking the improvement into account, one's overall success will still be less than that of the expert. The crucial step of Raz's argument is the claim that *only if* one follows the expert in *all cases*, so that his advise is taken as a reason for action that *excludes all the underlying reasons* on the basis of which the expert was supposed to

[287] Raz (1986: 67-69)

reach its judgement, would one have a success rate equal to that of the expert. Thus *maximally* improving conformity to reasons warrants the Preemption thesis.[288]

Is this account of a successful decision strategy a good response to the problem of instability? Apart from the problems with the rationality of following this strategy in situations of uncertainty (its main sphere of application), demonstrated by Scott Shapiro,[289] there is a futher and very important problem. The problem is that it has no resources to provide a boost for the plausibility of the Preemption thesis as well. For it to be an argument for the Preemption thesis, it would need to justify following authority/the expert even in cases of the expert making clear mistakes. This is the test case, which Raz's solution fails here as well.

The maximising logic that drives the argument and presumably accounts for its success - simply improving conformity to reason may justify treating expert's utterance as advise, *maximizing* this conformity is what turns it into protected reason with exclusionary force, ultimately undermines it. The logic is again simple. When authority occasionally makes a clear mistake,[290] and one happens to notice it at no extra cost, one can only maximise his conformity to reason if he does not act on this particular mistaken directive (and acts on his judgement, though generally with a lower success rate than the expert's, on this occasion is certainly better: anything above 0 – this is a case of clear mistake, is better), but treats its directives as protected reasons for action on all other occasions. After all, one could thus do better even than the expert himself! If so, the maximising logic ultimately turns against the Preemption thesis. Thus the Razian objective of developing a stable decision-strategy, which would provide a boost for the preemption thesis, is not met by his argument.

Thus, it is critical for the success of a Razian solution to the instability problem (and for his defense of the Preemption thesis) that it be shown: such balancing not only is not required on instrumental grounds, but should not be allowed, because it is always counterproductive. What needs establishing is not only that such balancing is inefficient because it involves high decision-making costs, but also that it is instrumentally

[288] Thus Raz obviously gives a maximizing interpretation of NJT (recall the discussion in chapter 2).
[289] Shapiro (2002a: 420-423).
[290] Raz's example is adding up integers and getting result with a fraction. This might not be an interesting example, but note that we have a clear mistake also in the stock example of the red light in the desert, which is the test case for whether the Preemption thesis works.

irrational, because it blocks the maximally improved conformity to ex ante reasons authority could bring.

Note that such argument is meant to work only in the case of *a legitimate authority*: an authority, one has good reason to trust, because it has a greater chance of bringing improved conformity to reasons than is the chance of the individual if left alone. Raz's claim, accordingly, is not the implausible one, *that the great mistakes* of authority should always be disregarded. Rather, his point is that if one has a good reason to trust an authority that it will not regularly commit great mistakes, and thus could bring better conformity to ex ante reasons, because it is less likely to be mistaken than its subjects, one should always follow its directives.

The promise of a successful solution, then, would depend on establishing that the paranoiac demand always to balance the underlying reasons in deciding how to act will defeat any advantages which could presumably be gained not only through having authorities, but even through adopting serious[291] personal rules, or by forming relatively stable plans for the future.

The above discussion helps me locate more precisely the source of the instability problem, which is the main obstacle to its solution as well. This instability is the predicament of the rational deliberator in one of his traditional embodiments – what has been called the *myopic* chooser. He understands instrumental rationality according to the "standard view": as requiring one always to decide how to act on the balance of all the reasons available at the time the decision to act is made. As a result of such decision strategy, with its "standard" account of instrumental rationality, the rational chooser may end up being worse off through successive instrumentally rational choices. The culprit, it seems, is the "standard view" of instrumental rationality.

A related strong objection to this "standard" interpretation of the instrumental rationality as always requiring such balancing was mentioned above: it makes incomprehensible the success of any minimally complex individual or collective activity, which is extended over time. Were the standard account of rationality be correct, it would render irrational

[291] The distinction here is between adopting serious (or mandatory) rules and following rules of thumb only. The serious rules are opaque – they do not allow considering the underlying reasons when following them, while the rules of thumb do not exclude this. Alexander (1999: 43-44), Raz (2001a).

to plan in advance, since one would be required always to reconsider his plan in light of all the reasons available at the time when the decision to act is due. Co-ordinating any complex activity, individual or collective, requires having relatively settled plans,[292] or relatively fixed modules, which are not to be constantly reconsidered.

> "We plan to avoid deliberating from scratch on each occasion of choice. And we plan to ensure the effective co-ordination of our action, both for the efficient realization of our own varied objectives, and for the fuller realization of those objectives through interaction with others." Gauthier (1996: 218)

Such considerations triggered developing more or less "revisionist" alternatives to this standard view of instrumental rationality. They may hold the promise of solving the "instability problem" that haunts the instrumentally justified decision to follow authority as well. But can Raz's account of authority employ them?

3. The Stable Instrumentally Justified Decision-making Strategy: Resolute or Modified Sophistication Choice?

The suggestion then is, that Raz's protected reasons with exclusionary force are possibly best interpreted as the above-mentioned relatively "stable plans" or relatively "fixed modules." Two alternative decision-making strategies, allowing one rationally to have such relatively fixed modules were offered. The first is the resolute strategy, developed by David Gauthier and Edward McClennen, which employs a thoroughly "revisionist" account of instrumental rationality. It was this strategy, and this revisionist account of rationality, which was criticized by Gregory Kavka, among others, through the Toxin Puzzle case. We have seen already, that the Toxin puzzle case is not analogous to the case of deciding to follow authority on Raz's theory and I take this to be an indirect indication that the optimal decision-making strategy in the case of Razian authority is not that of the resolute chooser. Michael Bratman's suggestion is that a better rational strategy for dynamic choice, which allows for having some such "fixed modules", or stable commitments is available, which nevertheless remains within the framework of the "standard" account of instrumental rationality. This is the *modified sophistication*

[292] This is the gist of Bratman's theory of intention as allowing to co-ordinate both intra-personal (individual) and interpersonal action over time.

strategy, which allows for having the benefits of reasoning with relatively "fixed modules" without committing the mistakes made explicit by the Toxin Puzzle case, but without also being caught between the Scylla and Harybdis of the "instability problem". Could Raz's account of authority be best interpreted as making use of this decision-making strategy instead?

3.1. The Resolute Strategy of Rational Choice

The backward induction argument I considered in the previous chapter demonstrated the predicament of what was called[293] the "sophisticated" planner. Both the sophisticated planner and the myopic chooser work with the standard account of instrumental rationality: but while the latter through adopting a plan ends up being worse off than he would be were he not to adopt any plan, the former escapes this predicament by not adopting the plan in the first place. Knowing that he will not be able to follow through with a plan, because at the time of the decision whether to follow through (due either to the presence of autonomous benefits, or to the absence of future benefits) he knows he will not have sufficient reason to act as planned, the sophisticated chooser cannot rationally adopt such a plan. Accordingly, he will not get the benefit of being helped in a reciprocation case, nor will he get the benefit in terms of improved conformity to reason in the case of rationally deciding to follow authority as well. The reason is that he is aware of the irrationality, according to his favoured account of instrumental rationality, of sticking to such a plan.

It was argued[294] that a stable rational strategy is available, which allows to get the benefits in TP type and reciprocation-type cases. This rational strategy is the already mentioned "resolute" strategy for dynamic (temporally extended) choice, developed by David Gauthier and Edward McClennen.

It recommends:

1. Adopting at T1 a plan for a course of action, which is best in prospect, given C.

2. At T2, given C, it requires following through with it, even in cases of changed preference-ordering/evaluation of the situation.

[293] The sophistication strategy was discussed and criticised by McClennen (1990), among others.
[294] David Gauthier (1996;1998 (b)). McClennen himself is committed to this strategy.

The resolute strategy privileges the preference ranking at T1of the alternative courses of action available then Adopting the prospectively best plan gives *a conclusive reason* to follow through at T2, given C, which trumps the reasons against following through, stemming from the changed preference ordering at T2. In cases of no unanticipated information (i.e. C obtains), the claim is, it is rational to preserve the intention to follow through and it is rational to act on it. This is rational, because if one know that one will persevere in one's intention and will follow through with the plan, even in cases of autonomous benefits, this will allow the resolute planner to get these benefits, by rationally forming the required intention (adopting the plan). Adopting the plan/forming the intention on the resolute strategy amounts to making a *strong commitment* to follow through with the plan/execute the intention, and undertaking such commitment is instrumentally rational, because it guarantees receiving the autonomous benefits. This account of strong commitments, let me stress again, violates the postulates of the "standard" view of instrumental rationality, requiring balancing of all the reasons available at the time of decision, which are still in the causal control of the agent in the time of deciding how to act: backward-looking considerations should not affect the rationality of a decision. The resolute choice model uses a revisionist account of rationality: it takes the plan-provided reasons as conclusive whenever instrumentally beneficial, even in the presence of autonomous benefits: i.e., even when they are essentially backward-looking.

The characterization of plans as providing trumping (conclusive) reasons could be taken to support the claim that Raz's protected reasons with exclusionary force account of authority is best understood as using this resolute choice decision-making strategy. [295]

The main charges against Raz's conception of authority could then be understood as being analogous with the charges against this strategy. But were this to be so, Kavka's Toxin puzzle case could be applied to Raz's analysis as well, since Kavka's TP was offered as a reductio ad absurdum of the so described resolute strategy. The fact that Kavka's TP case is not analogous to the case of deciding to follow authority shows, I believe, that Raz's account of authority does not use this decision-making strategy. Let

[295]"If I have adopted a plan and am reasonably not reconsidering it, then I have plan-based reasons for restricting my deliberation to actions compatible with my plan." Gauthier (1996: 221)

me, nevertheless, rehearse the main problems with it, revealed by the TP case, because this will help me in the discussion of the alternative, modified sophistication strategy, in the next section.

TP challenges the resolute strategy's "revisionist" view of instrumental rationality. According to this view, in deciding how to act at T2, it is rationally required to evaluate *courses of action as a whole*, even if parts of those courses of action are antecedent to that decision (and thus are not anymore in the causal control of the agent). To require evaluating *whole courses of action*, even when parts of them are already in the past, is a consequence of privileging the evaluative ranking at the time of the plan-formation T1 over the evaluative ranking at the time of the decision to act (plan execution) at T2. The debate between the sophistication - accepting the standard view of instrumental rationality, and the resolute choice theorists - substituting it with a revisionist view, is whether something - that part of a whole course of action, involving receiving the benefit, which justifies adopting the course of action in the first place, - though not anymore in the causal control of the agent, can nevertheless instrumentally[296] justify his action.

In short, the debate is whether *instrumentally rational commitments, which involve autonomous benefits,* are possible. By showing that it will be impossible to rationally form an intention to follow through with a plan, if the justification for forming the intention is autonomous with respect to executing that intention, Kavka has raised serious doubts concerning the rationality of the resolute strategy. The challenge is against the coherence of the revisionist account of instrumental rationality employed by that strategy. Raz's conception of authority, as already argued, does not employ the resolute strategy of dynamic choice. We are, then, left with the question what decision-making strategy would allow a rational planner to get the instrumental benefit of following authority, by providing for relatively "fixed modules," functionally equivalent to the authoritative protected reasons, if adopting the resolute strategy is a non-starter?

[296]The debate is about *instrumenta*l rationality and not about whether one may be *morally required* to fulfil a promise, stick to an agreement, or commitment to another person. Other, non instrumental justifications for the normativity of such practices are readily available: they are not my concern here.

3.2. Bratman's Modified Sophistication Strategy - Apt Strategy for Raz's Account of Authority?

3.2.1 Bratman's Strategy

It is worth then exploring the possibility of employing a modification of the sophistication choice strategy, offered by Bratman,[297] for the purposes of Raz's account of authority.

On Bratman's version of the sophisticated strategy, one will be able to get the benefits of sticking to some commitments (avoid temptations, for example, thus avoiding the "instability problem") though the benefits such as those in the Toxin and the reciprocation case will still not be available. The advantage of Bratman's solution is that it explains how stable commitments are possible without abandoning the framework of the "standard" account of instrumental rationality. This part of my chapter, then, will revolve around the issue whether the modified sophistication strategy offered by Bratman can account for how it may be rational to treat the authoritative directives as protected reasons with exclusionary force.

As a preliminary step, Bratman shows that employing the unmodified sophistication strategy will prevent one from getting even non autonomous, future-oriented benefits. One may at T1, for example, predict, that faced with a temptation at T2, he will find it difficult to stick to a plan, adopted at T1, and defect, thus leaving one at the end being worse off than he would be, were he not to adopt the plan in the first place. Because this advises against adopting plans in cases of temptation, the sophisticated chooser will be prevented from getting even future-oriented benefits.

The analogy with the authority is straightforward. One is rationally advised not to decide to follow authority, if one knows that every time one is faced with a case of disagreement with authority, one will fail to comply with its directives, thus defeating the point of deciding to follow it, and consequently, ending up being worse off than he would be, were he not to decide to follow the authority in the first place, and act on one's own judgement instead.

Bratman's solution explains how one may rationally forego giving in to temptations, without abandoning the standard account of instrumental rationality. The way he achieves

[297] Bratman (1998)

this is by introducing the concept of anticipated future regret. Thus the picture of the dynamic, inter-temporal choice is further enriched to include the perspective of the agent at T3, which temporally succeeds the decision to act at T2. Thus, in deciding what is rational to do at T2, one does not only evaluate the alternatives available then, based on the agent's preference ordering at T2, but requires that the decision is made in the light of the agent's preference-ordering at T3 instead. Since T3 is still in the causal control of the agent, the standard account of instrumental rationality is not altogether abandoned, but merely modified: the decision whether to stick to the plan is based not on the reasons at T2, but on the reasons that come from one's expected preferences in T3. Thus, if the agent expects that he would at T3 regret one's choice at T2, that choice is not rational and should not be made.

Thus Bratman's solution is to suggest to privilege still a different evaluation of alternatives than the one present at the time of the decision to act at T2. The arguments against privileging the T1 ex ante evaluative ranking (which was the resolute strategy's solution to cases of instability and temptation), however, do not affect Bratman's position, since the evaluative ranking at T3 is based on evaluating courses of action still available at T2 (the time of the decision to act) and can thus justify acting as decided (because getting the benefit is not causally independent from the action). Next, the justification for privileging the ex post T3 evaluative ranking depends on recognizing that a case of temptation is involved: the evaluative ranking at T2 is seen as a case of *temporary reversal* of an otherwise *stable* preference ordering. The agent firmly *identifies* with his preference orderings at T1 and T3. This is what justifies adopting a plan/deciding to stick to one's preferences at T1. Note, however, that the reason for actually sticking with it at T2 is *not* simply that one *has decided to do so at T1*. This decision itself here does not provide conclusive reason for action (as it does on the resolute strategy). Rather, the justification for acting as decided is that one anticipates that one will at T3 regret abandoning one's initial (T1) decision at T2. The decisive consideration here is not backward-looking, but forward-looking (one's regret at T3), and this allows the modified sophisticated strategy to avoid the serious problems with the resolute strategy's revisionist account of rationality.

3.2.2. Applying Bratman's Strategy to Razian Authority?

There are, I believe, serious problems with applying this otherwise plausible strategy to Raz's protected reasons account of authority, and they are manifold.

Firstly, Raz's account of when it is rational to decide to follow authority relies on establishing that by doing so, one will better conform to one's ex ante objective reasons, and not to one's subjective preference ranking. For Raz, one has an objective reason to achieve better conformity to reasons even if one does not actually prefer it. What is privileged in Raz's view is not the subject's evaluative ranking at T1 and T3, with which one identifies but rather the perspective of the correct balance of objective reasons. It is not immediately clear where to locate this perspective: is it T1, or T2, or, finally, T3? The fact of the identification with a perspective does nothing to clarify this issue.

Secondly, the case of the disagreement with an authority is not best seen as a case of temptation, produced by the temporal reversal of one's subjective preference ordering. The disagreement is rather about what better conformity to objective reasons requires on a particular occasion. When seriously disagreeing at T2, one does not expect that one will change one's mind at T3. Even less does one agree to take the perspectives of T1 or T3 as critical, simply because one identifies with them. If one seriously disagrees, this is done on the basis of one's convictions, and it is plausible to assume that one identifies with one's convictions. So, in case of serious disagreement with authority at T2, one cannot decide to follow the authority, simply because one firmly identifies with one's decision at T1 and with one's anticipated regret at T3. If there is a T, with which the agent identifies in seriously disagreeing with authority, it is T2.

Third, even granting that one need not necessarily identify with one's present convictions, on the basis of which one disagrees with the authority, it will still *be problematic to fix an exact point in the future in a determinate way*, so that at it the then-present-convictions will serve as the basis for criticizing the convictions at T2.

It might be claimed, for example, that at some point Tn, one may be convinced that disagreeing with authority was not reasonable, since by following the authority on all occasions (both when one agreed and when one disagreed with it) one actually would have done better than he would have done, had he acted on one's convictions when disagreeing with authority, and followed it on all other occasions. The problem is how to

"fix" in a determinate way this point, so that the anticipated future regret (providing the reason to stick with the decision) can have determinate value. The decision to follow a particular authority at T1 was justified on the basis that some benefits are reasonably expected from following that authority. At T2 it will be rational to follow through with the plan/stick with that decision only if some further overall benefits are expected. This implies that at T2 one will need to reconsider whether it is still rational to stick with the decision. All the benefits received up to this point are to be subtracted from the benefits that were initially expected, and it is to be calculated whether the future expected benefits still justify following authority, because one will still thus do better in conforming to one's reasons. (If the benefits that made following authority overall rational are already received, the decision to follow through with the initial decision will be irrational: this was the lesson of the TP).

Doing this calculation presupposes that one has a way of reliably knowing how the benefits of following authority are distributed in time, and there are problems with this assumption. If one had a way of knowing this, so that one could identify when following authority brings benefits, and when not, one would not need to follow authority in cases when one knows that following it does not bring benefits. The jurisdiction of the authority would accordingly shrink to cover only the cases when one expects following authority to bring benefits. One's uncertainty concerning the distribution of the benefits in time is one main consideration, which justifies taking the authoritative directives as protected reasons with exclusionary force, not to be generally balanced against the reasons subjects have independently of authority. Under such conditions of uncertainty, then, one will not be in a position to make the necessary recalculation of the balance of benefits of following authority. But if one does not recalculate the balance, one might be criticized for acting irrationally, since one's action of following the authority may not causally contribute to actually producing the benefit of improved conformity to reasons (in the case when this benefit is already in the past).

Let me nevertheless grant, for the sake of the argument, that it is somehow possible to rationally do such recalculation. Then Bratman's suggestion is that if one expects to *regret* (on the basis of this re-balancing) giving up the decision to follow authority, one needs to stick with it. There are problems with this suggestion as well, however.

The *fourth* problem stems from these difficulties with fixing the relevant future point of reference for determining the value of the anticipated regret. Decisions to follow authority rely on one's expectations concerning the overall benefit of following authority. Specifying these expectations relies on being able to set a determinate temporal limit on the expected interaction with the authority. However, this limit on the interaction is not independent of one's expectations for benefits. To the extent it is up to the subject to decide whether to continue to follow authority (the authority might for some reason stop making claims to obedience, and then it will not be up to the subject to decide whether or not to follow that authority), the interaction will continue till it is expected to be beneficial. The expectation, justifying fixing the terminating point, will, however, be constantly modified during the interaction with authority. This is so, because this interaction is a source of important information concerning the reasonableness of such expectation. But if this is so, one cannot have antecedently fixed point, at the end of which the balance of reasons is made and a verdict is reached whether one should/should not regret following/not following through with one's decision to obey authority. Without fixing such a point, the anticipated future regret will lack determinate value, and accordingly, cannot play the role of action-justifier.

The above argument, if sound, enables me to discern a further problem with Raz's approach of establishing the rationality of deciding to follow authority. On the one hand, this rationality depends on one's having reliable expectations that the authority will bring certain benefits in the future. It thus presumably depends on the extent to which one has reason to trust the authority. It is ultimately a matter of having reasonable beliefs and of acting in accordance with them.

On the other hand, rationality is understood to be a matter of objectively improved conformity to one's objective reasons. Thus what rationally justifies *following* the authority is that it *actually* brings improved conformity to reason. If this second view is right, the rationality of *deciding to follow through* should depend on whether the authority actually, as a matter of fact, brings such conformity to reason, and not on whether one has reason to expect that this will be so. One's own expectations (beliefs) even if reasonable, do not make it rational to follow authority, in cases when following

authority turns out not to be supported by reason (since it does not in fact bring improved conformity to reason).

This problem is familiar from the first part of this thesis, where NJT's interpretations were discussed. I argued there that an objectivist interpretation is more faithful to Raz's view on justification, though Raz recently admitted the need for a subjective element: he agrees one has to have a (subjective) reason to find out whether the putative authority has a reasonable prospect of being beneficial for him. Raz does not seem to have abandoned the objectivist interpretation, and for good reasons (then his account of authority would be dangerously close to describing authority as theoretical only). However, the problem I pointed out, remains. If Raz's considered position is that what makes rational deciding to follow authority is that it does, as a matter of fact, bring improved conformity to ex ante reasons, one might turn out not to be able rationally to decide to enter in interaction with authority, since one can only rely on one's expectations in making this decision.

Let me try to be as clear as possible on this last point. The problem stems from Raz's objectivist interpretation on when following authority is rational: namely, if and only if it brings improved conformity to ex ante objective reasons.

1. Start with an *objectivist* view when following authority is rational: IFF it brings improved conformity to ex ante reasons

2. Determining this is possible, if at all, only ex post, from the perspective at the end of the interaction with the authority. This is so, because NJT is cumulative, not a one-shot test of justification; so we will have to evaluate in the aggregate, at the end of the interaction, whether we have benefited from the exchange with the authority.

3. Determining when exactly that end is to be fixed, however, is done on the basis of the *subjective expectations* concerning the potential benefits from following authority.

4. Thus the rationality of following authority (on the objective reading of improved conformity) will, to an extent, depend on one's luck in having such expectations, which will as a matter of fact fix the end of the interaction in a way, that the balance of reasons for following authority turns out to be positive.

5. This argues, I think, for very short terms of interaction, so that one can be maximally in control of fixing the end of the interaction so that it turns out positive overall.

6. At the limit, it will be rational to have an overlap between the expectation and the calculation of the benefits at the end of the interaction, so that no discrepancy between the expectation and the actual balance of reasons is allowed for. This means one is rationally required to calculate the balance of reasons for and against following the authority on each occasion. We have abandoned the cumulative reading of NJT – one-shot interactions seem the rational thing to do.
7. This will clearly put an end to the possibility of having any beneficial interaction with authority.

We are back with the problem of rationally deciding to follow authority. Because of the inherent instability of the instrumentally justified strategy, it might be irrational to decide to follow authority since one may end up being worse off than he would be, if he did not so decide in the first place.
My conclusion of this section is that Raz's account of when it is rational to decide to follow authority cannot use Bratman's modified sophistication strategy for dynamic rational choice. The main problem is that on Raz's account it is not possible to fix in a determinate way a point in time, in reference to which the purported "decision-justifier" - the anticipated future regret, can have a determinate value. Far from providing solution to the instability problem, that would boost the Preemption thesis, the discussion of Bratman's rational decision-strategy for temporally extended choice allowed us to see a further problem with Raz's position. The discrepancy between the ex ante evaluation of the decision (in terms of expectations) to follow authority, and the ex post evaluation of the act of following (in terms of the actual balance of reasons), where the ex post perspective is privileged by Raz, argues against the rationality of deciding to follow authority.

4. Beyond the Decision Model: A Critique of Scott Shapiro's Constraint Model of Authority

The discussion so far has shown, I believe, the strains of providing a stable instrumentally justified decision-strategy of always following authority. The problems both from the standard view of instrumental rationality (instability) and from the revisionist one

(irrationality because of presence of autonomous benefits) may be taken to warrant a radical departure from what is a common presupposition for those two models. The common point is that compliance with authority (following authority on a particular occasion as it asks and because it asks to be followed) is considered a matter of decision on the part of the subject whether to comply. This is the position of both Raz and his critics. Scott J. Shapiro proposes to break free from this presupposition as the only way of overcoming the problems with the instrumental justification of authority:

> "The mistake made by all of these accounts of authority [Raz's and his critics'] is their assumption that willing obedience to authoritative directives is the outcome of some form of decision-making." Shapiro (2002a: 415)

He proposes an alternative, "Constraint" model of authority, which still works within the framework of instrumental rationality. It is important to see whether this model succeeds, before considering the possibility of abandoning (or at least supplementing) the instrumental justification of authority with a wholly non-instrumental one (or such elements).

The constraint model presents authoritative directives as "instrumentally valuable, when, and only when, they are capable of affecting the feasibility of non-conformity."[298] The suggestion here is that deciding to follow authority involves trying to do to oneself internally what Ulysses managed to do externally: bind oneself.

> "It is to forego later choice by the operation of the will, but it is as real as using some pre-commitment mechanism." Shapiro (2002a: 418).

When successful,[299] submission to authority makes any practical decision-making about compliance with concrete authoritative directives irrelevant: in submitting, one's present self binds, or causally constrains one's future self, by ruling out the possibility of non-compliance with authoritative directives.

---

[298] Shapiro (2002a)

[299] "The Constraint Model deals only with successful submission, where the agent actually follows through on the directives issued." Shapiro (2002a: 418)

This model has considerable advantages. Firstly, the problem of instability of the instrumentally justified strategy to follow authority obviously does not affect it, since it is not presumably a matter of decision whether or not to comply on a particular occasion. Secondly, the problems with the irrationality[300] of following clearly or greatly wrong authoritative directives also do not arise, since it is again not a matter of decision whether to comply or not. One cannot choose or decide not to comply, since non-compliance is not within the feasible set of options, once one's commitment to authority is successful (has taken effect). Thirdly, notice also that these advantages are gained without obviously embracing the resolute strategy either. Though Shapiro's solution resembles this strategy in that it privileges the present self, which effectively constrains its future self, it differs from it precisely in that on his account it is not a matter of decision on the part of the future self whether to stick to the present self's prior plan or decision. The future self has no choice in this regard: its hands are tied. Accordingly, the Toxin Puzzle's case reductio ad absurdum of the resolute strategy does not affect Shapiro's model: because it is not a matter of decision whether to follow through with one's decision to always follow authority even if no further benefits are expected (even if the benefit of authority is autonomous), one's present self has no problem of rationally forming the intention that his future self follow through with the plan.

Let me, nevertheless, marshal a couple of arguments against this model, which raise doubts concerning its success in addressing the problems from the instrumental justification of authority.

This model attempts to present authoritative directives as reasons for action with unwavering binding force. If commitment to authority is successful, if one's present self has managed to constrain his future self to act in accordance with the demands of authority, the future self becomes unable to act contrary to the will of the authority. In this, the model provides a good account of the binding force of authoritative directives. To the ready objections about the irrationality of following such rigid authority that can in principle commit both clear and great mistakes, and still demand our absolute

[300] "Compliance with otherwise believed wrong directive is the only feasible, and hence the only optimal action, when an agent is successfully committed to authority deemed beneficial." Shapiro (2002a: 419)

obedience, it flatly replies that when successfully submitted to the authority, we have no choice in this regard but obey.

Precisely at this point of apparent strength I see the weakness of this model. It seems that my reason for obeying authority is not that *authority so requires*, but, rather, that *I have committed myself* to such obedience. If this is so, the model will not provide a good account of practical authority: practical authority it itself presumed to create new reasons for action of a special kind – both CiR and ER.

Let me illustrate the point using Shapiro's example of two obese friends (Larry and Charlie) and their trainer Sonny.[301] Shapiro introduces it (both friends prefer to exercise, but one of them only needs an authority to help him act on his preference) in order to explain what is missing from the "decision" models' explanation of the normativity of authoritative directives. What is missing is the essentially "volitional" element of those directives. For Shapiro, the directives are not tools for making decisions, but (external) ways of preventing decisions from being made."[302] Thus, Sony's order to Larry to go and exercise is not an input in Larry's own deliberative process, but rather a causal constraint on Larry's non-conforming to that order behaviour.

My contention with this example, and the model which it illustrates, is that this causal constraint is only effective, and thus authority's *will* has relevance to how one acts, only to the extent and till one remains committed to obeying its orders. *One's obedience to authority is conditional on one's own commitment to that authority.* But if it is a matter of subject's own commitment to authority whether or not one complies with the authoritative commands, it remains an open question whether and what exactly Larry stands to gain from the services of Sony as an authority. On this description,[303] there seem to be little (if any) relevant difference between the case of submission to an authority and the case of intra-personal commitment. In both, one could 1. exhibit a weakness of will, 2. change one's mind in light of more information (that the directive is

[301] Shapiro (2002a: 416-417)
[302] Shapiro (2002a: 418)
[303] The story may be made more complex, by adding details about the relation between the subject Larry and his authority Sony. It might be, that Larry's duty to obey Sony, though initially started as conditional on the presence of the commitment of Larry's later self to his previous self (which has decided to submit to the authority of Sony), has developed and is self-sustaining, probably even self-validating. I do not dispute the possibility or the plausibility of such a scenario, but this is not the description Shapiro offers, nor will it correspond to the points he wants to make against the Decision model of authoritative directives.

mistaken, say), or, 3. simply change one's preferences, where all the three can lead to abandoning by one's future self the commitment of one's present self.

My conclusion here is that if one's compliance to authoritative directives is conditional on one's upholding one's commitment to authority, it is not clear what, if anything, authority adds to one's commitment, and whether it makes a practical difference to how one should act. The practical position of the two friends (one intra-personally committed, the other – inter-personally committed to obey an authority, where obeying authority is nevertheless conditional on his continuously upholding his own commitment) seems structurally and practically indistinguishable. The constraint model seems reduced to the resolute choice decision model, and could be subjected to all the critiques against the latter.

Further, note that the problems, which motivated Shapiro's general critique of the decision model is replicated in his Constraint model as well. The main problem, coming from the instrumental justification of authority, is that if obedience to authority on each occasion is conditional on authority's success in getting the balance of ex ante reasons correctly, this defeats its "authoritative" character. True, on the constraint model, obedience to authority is not conditional on its success directly. Nevertheless, obedience is conditional on one's remaining committed to the authority. However, even if one cannot directly ask whether it is instrumentally justified to obey a given authoritative directive, one could (indeed must!) always ask whether it is rational to remain committed to an authority which, as it may well happen on a particular occasion, issues a grossly sub-optimal directive, and this is apparent to its putative subject. The rationality of remaining committed to authority in such circumstances is questionable: rationality may in fact require reconsidering one's commitment instead. But if this is so, the rationality of remaining committed to authority, on which one's obedience to its directives is conditional, does again depend on authority's success, if only indirectly so. The problem of the instrumental justification of authority reappears at the level of the rationality of remaining committed to a failed authority.

The explanation for this reappearance of the problem, I believe, is that while Shapiro needs some causal mechanism, which would guarantee that disobeying authority is not

feasible, the devise he uses – internal commitment, as shown above, seems too feeble to provide such a mechanism.

Notice that he cannot go for the stronger devise – say, external pre-commitment of the type Ulysses resorted to, setting in motion some causal process in the world, which would cause one to act as commanded (ex. Making sure that one will be locked in a room as a measure against disobeying the order to stay in the room). Such an external causal mechanism would only guarantee conformity to, and not compliance with authoritative directives. This is so because external causal mechanisms do not simply "repress" and thus remove the reasons we have for disobeying a directive from the feasible set of options, leaving it to us to act for the reason that the authority so commands. Rather, they directly "make us" do as the authority commands, even against our considered judgements that the reasons against obeying are weightier. Such an externally caused action would not be an intentional action – one would not have acted for a reason at all, and certainly not for the reason that one has been ordered so by an authority. This is why the "pre-commitment" strategy of rational dynamic choice is a non-starter as an explanation of rational rule-guided behaviour, and of being guided by an authority in particular. [304] But as we saw, the two "commitment" strategies – the constraint model and the Resolute choice model, also have their problems.

Shapiro's Constraint model of authority, an initially promising alternative to the Decision models of authority, fails to provide a solution to the rationality problems, coming from

[304] It may be objected this argument does not apply to precommitment in general, but to one type of it only, using external causal mechanisms specifically. This is not the only possible understanding of pre-commitment. For example, in the course of building his case for a substantive reading of constitutionalism and judicial review, using the concept of precommitment, Janos Kis (2003: 194-202) argues that it is possible, indeed rational to bind oneself by instead of setting a causal mechanism in the external world, one entrusts the decision to a person, in the expectation he will execute one's intention of reaching the best decision, supported by that person's considered judgement. There are problems, however, with construing this latter case as a case of pre-commitment, of true self-binding: in what sense is such self-binding truly irrevocable by the self-binding person himself? Thus, one has to distinguish *causal* from *normative* precommitment and describe the case of collective precommitment in constitutions as normative: constitutions confer rights and impose duties. Jon Elster's (1984; 2000) treatment of the "Ulysses analogy" for constitution adoption is criticized for failing to do so, in Shapiro (2002c: 178-181). Be this as it may, my point here is that to the extent the action of following the directives of the entrusted with the decision other person, could be attributed to our agent as an intentional action, it is action not based on reasons, but one *externally caused*. The fact that the external causal "mechanism" is a reasoning agency (which is indeed relevant in evaluating the rationality of authorising him, rather than employing, say, a guillotine) does not change the fact that our individual does not *follow*, obey the authority in the required sense, after pre-committing.

an instrumental justification of authority. If my observations are correct, this account comes dangerously close to, and probably collapses into the Resolute choice version of the Decision model with all the already discussed problems with it.

5. Conclusion

My conclusion is that both the Resolute and the Modified Sophistication strategies for dynamic choice, as well as the alternative Constraint model, fail to provide a solution to the rationality problem with deciding to follow authority. The problem is due, recall, to the inherent instability of the instrumentally justified strategy of following authority. These models either fail in their own terms, or as applied to Raz's account of authority as providing subjects with protected reasons for action with exclusionary force. This warranted a detailed discussion of the instrumental justification of authority on Raz's account. The conclusion reached is that the above problems could ultimately be traced back to the instrumental character of that justification.

Part Four

The Authority of a Liberal-Democratic Political Order

The analysis in the preceding parts of this thesis has been focused on important theoretical issues, concerning the coherence of Raz's model of practical authority and its capacity to solve the rationality paradox that has traditionally puzzled the students of authority. Many problems have been identified with this model in these respects. I have also analysed in detail Raz's conception of legitimacy, especially the adequacy of NJT as a legitimacy test, its relation to the Preemption thesis, etc. I have also argued for the validity of certain type of reasons – agent-relative reasons of autonomy and partiality, that may require imposing external limits on the exercise of authority and deflating the overarching claims to comprehensive supremacy over all other normative domains such authorities necessarily make on Raz's account of the political and legal authority. Liberal authorities might be characterised precisely by refraining to make such obviously implausible claims: if so, Raz's account of political and legal authority may need some revisions. Nothing has yet been said, however, concerning the issue whether Raz's conception can account for the authority of *democratic* political arrangements specifically. It is time to attempt to remedy this.

Chapter Eight

Content-independent Reasons for Action by Democratic Pedigree?

1. Introduction: Democratic Authority and the Content-independent Reasons Problem

Democracy is a philosophically puzzling concept. It requires that people accept the outcomes from certain specific procedure – notably majority vote – as authoritatively binding, regardless of the substantive merits of these outcomes. How could this be rational? And can democracy be authoritative, i.e. can it be legitimate in principle? Though our considered conviction is that employing certain democratic procedures for collective decision-making is a necessary if not sufficient component of the best form of institutional arrangement for modern societies, we are puzzled what accounts for their capacity to yield binding decisions given their peculiar features: the decisions resulting from the democratic procedures are deemed binding irrespective of their substantive merit.

One of the distinct features of the concept of authority employed by Joseph Raz's Service conception of legitimacy, is that authority, when legitimate, provides valid content-independent reasons to its subjects. Many find this concept no less (some find it more) puzzling than the concept of democracy itself: it validly requires from a person to do F, irrespective of the merits of F-ing. It is this latter concept, which may hold the promise of accounting for the specific, puzzling features of democratic decision-making: that the results of such a procedure are binding irrespective of their merit. The theoretical benefits are worth the effort: that, which is puzzling in one of the concept may help in finding a solution for the puzzles in the other. Thus the two puzzles may be easier to solve together. But can this be done? These are some of the questions I address in this concluding chapter of my thesis.

Can a case be made for the authority of political arrangements employing a democratic decision-making procedure on the ground that the results thus reached give valid content-independent reasons for compliance with them? Or may be a stronger argument could be advanced: the fact that democratic procedures could yield such valid CiRs (while others meet some difficulties in this respect) favour them over political arrangements not providing for such procedures?

These are the questions that drive my argument in this chapter. An inconclusive answer is provided at the end of a long, tortuous discussion. The framework of the question is Raz's account of practical authority: on it, authority's directives give subjects content-independent reasons for action (CiRs). Though some preliminary discussion as to what it means to have CiR was already offered in the first chapter of this thesis, a further analysis is required to establish whether this concept is coherent. Even if the answer to this conceptual issue is positive, one should further inquire whether there are such valid reasons, and whether and when is it rational to act on them. Only after this preliminary work is done, will I be able to address the question whether valid CiRs can be provided by an authority with democratic pedigree and whether it is rational to act on them. I show that a positive answer to this question is warranted.

On Raz's account of practical authority, authoritative commands provide agents with protected reasons for action (comprising CiR and ER): it defines authority as practical authority in the narrow sense. Alternative to this model accounts define authority as theoretical only, or as influential, the difference between all the three consisting in the different interpretations given to the authoritative utterances and the different types of reasons they give to the subjects of the authority.

The model of practical authority has been subjected to strong critiques on the ground that either there are no valid exclusionary (and by implication, no protected reasons as well) reasons, or that the very concept of exclusionary reason is incoherent, or both. There are critics, who accept that authoritative utterances give CiR - they opt for the model of influential authority. Some go further and contest the CiR character of authoritative reasons – they deny that authoritative utterances provide reasons for action, opting for a

model of authority, defining it as theoretical only, giving *reasons for belief* in the validity of some pre-existing reasons for action.

Here I address the second, more radical challenge against the coherence and the validity of CiR. My interest is not, however, in primarily defending the model of practical authority (as a plausible account of political authority) per se. Rather, I prepare the ground for testing the hypothesis that a plausible case for a political arrangement employing democratic decision-making procedures can be made, on the ground that the reasons for complying with the decisions reached in this way can be valid. Thus I focus on establishing the possibility of having valid *content-independent reasons* by democratic pedigree. Before that one is to show, however, that the very concept of CiR is not itself incoherent.

I build on the definition and the discussion of the problems with the coherence of this concept, provided in the first chapter of this thesis. I specifically focus on Raz's response to "the normative gap" problem with CiR's coherence: he follows an outcome-based, instrumentalist strategy. He introduces exclusionary reasons (ERs) as that element in autonomous reasons (comprising CiR as its other constitutive part), which brings a break in transitivity of justification, without thereby threatening the validity of the resulting reasons. I find faults with this Razian solution in terms of ERs, especially within an instrumentalist justificatory framework. Many of the arguments against the indirect instrumentalist solution to the paradox of rationality and authority, discussed in the third part of this thesis, are brought to bear here.

I then address the issue whether an argument for the existence of valid CiRs can be advanced, if an alternative - proceduralist and non instrumentalist, strategy is pursued. The hypothesis I test is that there are valid CiRs, having value not depending on the evaluative properties of the action they recommend, when these reasons have their source in an authority with a democratic form of decision-making. This argument has three steps. First, I demonstrate that even if the *value* of a procedure is what validates an authoritative reason, we still have a case of CiR. Next, I advance some arguments for maintaining the distinction procedural value/outcome value. Third, an argument from democracy for the existence of procedural merits/value is briefly considered.

If the arguments in the chapter are sound, democracy could indeed provide authoritative procedures of collective decision-making, yielding results that rightly claim legitimacy. What is more, the stress on procedures in democracy may turn out to be a serious advantage of this form of government as compared with other competitors. This line of arguments gives ground for favouring a democratic type of authority, since it shows that such authority could in principle provide valid CiRs. Some limitations of the above arguments, which give rise to problems for further research are outlined at the end.

2. CiRs: The Normative Gap Problem.

Raz's definition of CiR is:

(A) "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."[305]

In introducing the concept of CiR in the first chapter, following John Gardner, I offered an alternative requirement for CiR:

(B) CiR is a reason for action, which is not dependent on the evaluative properties of the content of the action, for which it is a reason

I have accepted the "positive" second part of Raz's definition of CiR as valuable and worth preserving, and have focused only on the diverging "negative" parts of these two definitions: whether only a direct connection with the required action is denied, or any dependence on its evaluative properties is denounced as well.

2.1. The Normative Gap Problem and the Service Conception of Legitimacy

Recall from the discussion in the first chapter of my thesis, that the most characteristic feature, as well as a central problem with CiRs is that they introduce what Raz calls a *normative gap* between what one ought to do (the normative force of the reason) and what is good about doing it (the value of the action). An explanation how despite the gap,

[305] Raz (1986: 35)

CiR can be in principle valid, is necessary in order to maintain that this concept is not incoherent.

The problem, recall, is that this normative gap is not local. It was this characteristic, I claimed, which demonstrates why the "no direct connection" requirement for CiRs is inadequate. It is insufficient to account for the specificity of CiRs, since it only denies connection between the CiR and the concrete action for which it is a reason. If the gap, however, is not local, then more is necessary than just denying the connection is direct. My further claim was that to be distinct, CiRs should not depend for their validity on the evaluative properties of the action they require. This feature, defining their distinctness, however, is responsible for the main problem with them as well. The charge is that such *reasons are incoherent*.

Those theorists, who share Raz's position about the dependence of normative validity on reasons, and insist at the same time that CiRs play crucial role in any account of authority, should be particularly concerned with this charge. The position that justification/validity is essentially or primarily based on evaluative considerations, coupled with the point that justification is in principle transitive, is threatened by admitting the validity of CiRs, since they introduce a normative gap, difficult to deal with within this theoretical framework. [306]

It better be possible to solve this problem, introduced by the presence of CiRs, and thus rebut the charge of incoherence, because it presents a particular problem for Raz's Service conception of legitimate authority, apart from the general puzzles it raises for the theory of practical reason/normativity. Notice that the normative gap will feature in any conception of authority, sharing Raz's account of the concept of authority, as providing CiR (and protected reasons for action more broadly) to its subjects, because of the general problems with this concept. The reason why solving it might meet *special difficulties* on Raz's conception of legitimate authority, is because the "normal justification" for the exercise of an authority on this conception is placed in the *good* of authority's serving its subjects (by producing good outcome in helping the subjects achieve improved conformity to their reasons). This position, conditioning the justification of authority on authority bringing about outcomes with certain evaluative

---

More on the normative gap problem, see section 2.2. in the first chapter of this thesis.

characteristics, does not sit comfortably with maintaining at the same time that an essential characteristic of any authority (a conceptual point about authority) is that it claims to provide its subjects with valid CiRs. The problem is that validity of such type of reasons is not conditional on such evaluative characteristics. It must, nevertheless, be possible to close this normative gap, opening with the presence of CiRs, so that the ultimate justification for an authority could in principle be in the good of having it. How can this be done?

2.2. Closing the Gap: Appeal to Merit at a Next-order Level of Justification?

One straightforward way of trying to do this will not do. On the "no direct connection" requirement for content-independence, the validity conditions for such reasons for action, can be put as follows:

'X has a valid CiR to *f* (to perform a particular act-token of the act type *F*), if *f*-ing is *not directly connected* to X's reason to *f*. The validity of X's reason to *f*, however, depends on X having a reason to *F* (perform any particular act-token of this act-type)'.

The explanation for the validity of X's reason to f here depends on X having a reason to F (perform any act-token of the act type F). This implies that the *force* of X's reason to F is "transmitted" to X's reason to f. X's justification to f is entailed by X's justification in acting on the reason to do the act-type *F*', which itself is entailed by X's justification to act as a meritorious authority commands, etc. This chain of justification-transfers shows, however, that here the transitivity of justification is not broken. If violating the transitivity of justification is a necessary condition[307] of content-independence (as it is for Raz), here we have no case of content-independence. If the transitivity of justification is nevertheless broken - then the validity of X's reason to f will not depend on X having a reason to F, but in some altogether "external" to both of these reasons fact: that f-ing has been commanded by Y, for example, - this account fails to show how it is not irrational to act on CiR. Either one is not justified to act on CiR, or if one is justified to act on the reason provided by a meritorious authority, this reason is not CiR, since the chain of content-dependent justifications is then not broken.

[307] Stronger: for Raz content-independence *means* that transitivity of justification does not hold.

2.3. Closing the Gap: Raz's Autonomous Reasons Solution

Raz offers an explanation of how CiRs can be valid, and the puzzle solved. It consists in showing that CiRs are constitutive part of autonomous reasons, [308] and acting on such reasons is justified and the reasons themselves – valid. Autonomous reasons (comprising both CiR and ER) are the reasons provided by promises, agreements, undertaking commitments, forming plans for future action, deciding to follow rules, etc. They are just a new name for our old acquaintance – the protected reason for action authority purports to create for its subjects. However, the new name stresses the importance of their content-independence component – they are reasons, autonomous with respect to the evaluative characteristics of the action they require. The protected reasons label Raz uses on other occasions, stresses more the characteristic feature of the exclusionary component instead – it *protects* the new reason authority provides by *excluding* acting on inimical to that new reason other reasons.

Autonomous reasons derive their *validity*, their force as reasons from the reasons, justifying giving a promise, making an agreement, committing oneself, etc., but they *do not depend on nor do they point to any good* or value in performing the action for which they are reasons. The force of reasons here does not go through. The reason for having a rule, making an agreement, pre-committing oneself, is not at the same time a reason (not, Raz says, under that description) for doing what the rule requests, the agreement demands, etc. Counterfactually, even though what the rule requests, the agreement demands, etc., could have been different, still the reason for adopting the rule, the agreement would have stayed the same. Thus a break in transitivity of justification occurs, testifying to the presence of content-independence.

One may deny that such reasons can be valid, since one may hold the view that validity of reasons and their normative force could only depend on the evaluative properties of the concrete actions for which they are reasons. Since autonomous reasons violate this requirement, it might be denied that there can in principle be such valid reasons. This is the incoherence charge. There cannot in principle be valid autonomous/CiRs, since the

[308] An extensive discussion of the problems with Raz's authoritative-as-autonomous-reasons-thesis was provided in the chapter on the Toxin Puzzle analogy.

normative force of the reasons necessarily depends on the evaluative characteristics of the action, for which they are reasons.

Raz's response to the incoherence charge is ingenious. If it is denied that there can in principle be valid autonomous reasons, it would be difficult to explain how making promises, agreements, commitments, etc., which create precisely such autonomous/CiRs, could make any difference to the reasons for action one has. Ultimately, it would be difficult to explain why the acts of promising, undertaking commitments, making agreements, are not altogether irrational: they would certainly be irrational, if the fact that one has performed such acts does not make any difference to the reasons one had prior to them. The explanation why these acts are not irrational points to the *value* of participating in such practices, of which performing such acts are constitutive. The *advantages* participation in such practices brings about make it rational to perform such acts, and also make it rational to act on the autonomous reasons[309] created by these acts. Thus, the simple point Raz makes is that unless we are willing to challenge the rationality of such wide-spread and considered legitimate practices as promise-giving, undertaking commitments, making decisions for the future, adopting rules, etc., we are not justified to challenge the coherence of autonomous/CiRs. Furthermore, if we are unwilling to challenge the rationality of such practices, we are not justified to be more suspicious of the content-independence of authoritative directives in particular (since both involve autonomous reasons with the same characteristics): they seem to stand or go together.

There is a strong reason to doubt the suggested by Raz symmetry between autonomous reasons by authority and autonomous reasons by promise, commitment, agreement, etc. Thus it might be that the concept of autonomous reason involved in the normative practices Raz lists, is a coherent concept, while the same concept is not involved in authority giving reasons for action to its subjects, or even if involved, other features of the latter case pose a threat to its coherent application there. This suggestion should not be immediately dismissed: there are clear differences between the reasons provided within these practices. The source of the reasons is different, and this may reverberate on their central features. The source of these reasons in the one case is the subject himself, in

[309] Raz's explanation why the reasons provided by the acts of promising, etc., are autonomous, consists in showing that only by taking them as autonomous (i.e. CiR and ER) one could get the benefits of the practice.

the other – the authority. Further, authority claims to provide its subjects with CiRs, changing subjects' normative situation (what they ought to do) with no other apparent justification than that the authority says so. The situation is different (one is tempted to say, radically different) in the case of an agent who by committing himself to a future action, adopting a rule, promising etc., creates autonomous reasons, which apply only to himself. The agent changes *one's own* normative situation by creating agent-relative reasons for action: they apply only to himself. Further, in the case of an individual agent creating autonomous reasons, he is "in full command of the nature of the act. Being the promisor, it is [he] who determines its content."[310] When authority purportedly creates autonomous reasons, in contrast, subjects owe it a duty to perform whatever acts authority commands simply "because […it] demands their performance."[311] One might not be willing to grant that authority can create valid autonomous reasons, because this amounts to "granting it the moral power to assign us particular duties, without retaining any control of their creation or contents."[312]

Admittedly, both authority and agents, creating autonomous reasons for its subjects only, or for themselves, respectively, pose a problem about the rationality of acting on those reasons. However, while there might be some explanation as to why it can be rational to form plans, make decisions etc. and act on them in the case of the agent, the same explanation will face much more problems in the case of authority. The explanation in the case of an individual agent either points to the function of the practices (of promising, pre-committing oneself, etc) in furthering agent's good (as he happens to understand it), and thus to their instrumental value, or to the constitutive for his agency value of having the capacity to create autonomous reasons (by willingly identifying oneself with a set of reasons, making them constitutive of one's identity as an agent). The problem with offering this type of explanation in the case of authority is that the authoritative reasons do not allow the individual subjects to retain control over the content and creation of reasons, they will be obligated to follow. Authority claims to give CiRs not to itself as a collective agent (in which case the above explanations might be good enough, though there are separate problems in the idea of collective agency itself as well), but to a

[310] Gans (1992: 98)
[311] Gans (1992: 99)
[312]Gans (1992: 99)

plurality of agents individually, without allowing them to retain control over these reasons. The threats involved in taking CiRs as valid and acting on them concern both one's rationality and one's autonomy in following authority.

In short, the problem with the analogy is that authoritative reasons do not allow the subject to *retain control* over the creation, the content and the scope of autonomous reasons. The power to change one's normative situation is granted to an authority, making legitimacy claims - to have a right to rule without recognising any external to this right limits. Giving promises, making decisions, etc., on the other hand, does allow the agent to retain control in this respect (one alone changes one's normative situation: the power to control it is not granted to anybody else). The analogy breaks further, when we recall that political authority and law in particular on Raz's account necessarily makes a claim to supremacy over all other normative domains, which further erodes the prospects of retaining control over one's normative situation, once one submits to such authority. [313]

This does not mean that there cannot be an explanation for the coherence of CiR and the autonomous reasons more broadly, which would permit authority to provide such reasons to its subjects. An argument for this is unlikely to successfully rely on the above analogy, however.

3. Types of Strategies for Validity of CiRs Distinguished

I concentrate now on the next problem: even if the normative gap problem (concerning the *coherence* of the concept of CiR) could in principle be solved by introducing autonomous reasons, it still remains to be shown that such reasons can be *valid*, and that it would be rational to act on them. The difficulties with establishing these two points (existence of valid autonomous reasons and rationality of acting on them) are obviously again related to the normative gap problem and the break of transitivity of justification CiRs bring in. Both validity of autonomous reasons and the rationality of acting on them

[313] I discussed the problems with these claims and their compatibility with Raz's Service conception of authority's legitimacy in the second part of my thesis. The discussion here, I believe, brings in further argument in this respect.

depend on closing the normative gap and establishing that despite the break in transitivity, one is nevertheless justified in acting on these reasons.

I suggest that we distinguish two main strategies of responding to the normative gap problem here: one is to try to solve it while remaining within a generally outcome-based, though not strictly speaking consequentialist[314] framework: this is Raz's outcome-based instrumentalist strategy. The way to show that there can be valid CiRs is to point to the advantages in terms of valuable outcomes – *improved* conformity to outcome-reasons, (though not necessarily to action-reasons)[315]– to be gained if one accepts them as valid. Outcome-based evaluative considerations ultimately back CiRs' validity, even though CiRs do not themselves contain direct appeal to evaluative considerations. Further, not simply valuable outcomes, but maximizing those valuable outcomes, is what backs this validity claim.

The other strategy is to leave behind this instrumentalist framework. The suggestion here is that in order to have valid CiRs, it is necessary to sever the connection between those reasons and *any outcome-based merits* (both of the directly commanded action itself and overall), with their inherently maximizing logic. This would require taking as relevant for determining the validity of CiRs only procedural merits.

A third approach is to embrace a dualist position and maintain that both outcome-based and procedure-based merits could validate CiRs. Such dualist position would hold that there are two distinct, irreducible dimensions of evaluation that may often come into conflict.[316] This position has its own problems, however, the main being - in what terms are these conflicts between the two dimensions of evaluation solved, if they are not

[314] Raz denies that his analysis of the concept of authority relies on any specific moral doctrine (consequentialist, deontological, or any other): as analysis of the concept that is shared by people favouring different moral doctrines, it should be compatible with all of them, Raz (1989: 1184 -85). Further, even when Raz discusses his own conception of legitimate authority, he still denies that it is consequentialist in character, strictly speaking (Raz 1986:chapters 11,12,13), since he rejects two consequentialist theses: that of commensurability and that of transparency of value.

[315] The problem with action-reasons was discussed in some detail in chapter 2 of this thesis.

[316] Such dualist position is defended by Christiano (2004): "I wish to defend an account of the authority of democracy that is holistic but that is not monistic. I shall call it a form of evaluative dualism with regard to the assessment of democratic institutions." Christiano (2004: 268) The two evaluative dimensions are outcome and procedure.

reducible to either of the two, nor an encompassing, middle term between them is provided.

Le me start with the monistic strategies – we have to show that at least on one of them CiRs make sense and could be in principle valid. If both succeed in this respect, we might need to face the dualist position and examine its success in surpassing the problem of conflict between the two orders of evaluation.

3.1. Raz's Instrumentalist Outcome-based Strategy

Start with Raz's strategy. The rationality of acting on CiR provided by promises, authority etc., is in the *value* of having such authority, having the institution of promising, etc. As noted, the value here is generally conceived in terms of "outcomes": *improved* conformity to reasons is the valuable outcome that ultimately justifies following authority, keeping promises, etc. Thus for Raz CiRs' validity will ultimately depend on a valuable outcome (measured in terms of improved conformity to reasons) with which a closure to the normative gap is provided. What backs the validity claim here and thus closes the normative gap is not simply "value" but "*more value*": if acting on CiRs would only bring as much value as acting on the first-order reasons directly, there might be nothing to commend acting on the former. An appeal to a *maximizing conception of outcome-based value* explains how CiRs can be valid on the Razian instrumentalist strategy.

How rationality of acting on CiR is preserved even though the transitivity of justification is violated (so that we do indeed have a case of CiR) is explained by introducing exclusionary reasons (ER). It is through showing how ERs can be valid and one - rationally justified to act on them, that Raz explains how one can act rationally on the CiRs provided by authoritative directives, even though the transitivity of the justification for one's actions is violated. Thus the focus in the Razian strategy is shifted from CiRs to ERs: demonstrating the validity and rationality of acting on ERs will demonstrate the validity and rationality of acting on CiRs as well.

3.1.1. Problems with Validity of ERs

ERs are peculiar. These peculiarities are warranted, and the reasons with such characteristics - valid, and acting on them – justified, since they are instrumental for achieving improved conformity to one's own reasons.

ERs thesis says: even if one is justified on the balance of first-order reasons to do A, one may have a valid ER not to do A *for the reason that it reflects the right balance* of reasons. It is thus perfectly alright if one does in fact do A, but only if one does it *for other reason* than that it reflects the right balance of reasons. For example, it is alright if one does A, but only if one does it *because* doing A was commanded by a legitimate authority (one that brings improved conformity to those same first-order reasons).

Is there a case here of intransitivity of justification?

The answer is yes. This is demonstrated by the counterfactual: it would have been perfectly alright if one did not do A (even though it was supported by the right balance of first-order reasons) if legitimate authority did not command doing A but doing B instead. Stronger than that: it would have been *wrong* to do A (even though doing A was supported by the right balance of pre-existent reasons, but acting on the balance of pre-existent reasons was excluded by the ER provided by the command) if authority commanded doing B instead.

Is it, however, indeed rational to accept ER as valid and act on it? How is this rationality established?

As in the case of any action, this rationality is established by appeal to the first-order reasons for action, applying to the agent. Here they are taken, however, in a wider perspective: one overall, though not in each separate case, conforms better to them. The reasoning is as follows:

a) following authority can bring improved conformity to one's first-order reasons for action overall because authority has an expertise and a capacity to solve co-ordination and Prisoners' dilemma collective action problems, and can economize on agents' decision-making costs (in terms of time and labour).

b) if it is rational to act so that one's conformity to reasons is improved, it is also rational to follow authority if thus one's conformity to reasons is indeed improved.

c) following authority can bring such improved conformity, *only if* one takes authoritative directives as providing him with valid ERs (as well as CiRs), even when one 1)disagrees with them and 2) is right to disagree with them, since authority did *not* in fact get the balance of first-order reasons right on this occasion.

Then, from a) +b) +c), follows:

d) It is rational to accept ERs as valid and act on them, even in cases when authority's command did not get the balance of pre-existent first-order reasons right.

From the numerous critiques against this strategy, the one I develop here takes issue with Raz's claim that rationality requires obedience to authority even when one strongly disagrees with it - whenever, despite his disagreement, his obedience is indirectly beneficial for the subject himself. Since rationality is measured in terms of improved prospect for good outcomes, it is rationally permissible to act against one's judgement if thereby one's prospects are improved. Further, authority does not preempt forming judgement. It only preempts acting on the result of this judgement, since acting as authority commands is what improves those prospects.

This sounds innocuous: authority only excludes directly acting on the balance of reasons, not deliberating on it. Subjects thus retain their status of independent rational agents, presumably threatened, if "surrender of judgement" was instead involved. It is thus rational to act on the exclusionary reasons authority provides.

Notice, however, a problem with the implications of this position: it is far from being innocuous. If authority excludes only acting on one's judgement on the merits of the case, and not forming the judgement itself (i.e. authority is practically exclusionary), this means that when authority commands action with which one disagrees (i.e. he has formed judgement on its merits), the intention one forms to perform the obedient action is in contradiction with the content of this judgement. Congruity between one's practical judgement and one's intentions is a central requirement of practical rationality,[317]

[317]The point, recall from section 2.6. of my second chapter, is that the final aim /end of practical deliberation is forming an intention to act, and it is a postulate of practical rationality that one's intention coheres with the process of deliberation that yielded it. May be Raz wants to say that authority permits one to form a non-practical, theoretical judgement, and it need not result in forming an intention to act, just in forming a belief how he should act. But then again, practical rationality does not seem to permit forming intentions in contradiction with one's properly formed, reasonable beliefs.

violated in the case of obeying authority. If so, when authority addresses the subject with a directive he disagrees with, the subject cannot rationally act on it, since there will be incongruity between what he judges he should do (believing it right), and what he forms an intention to do. The problem is that performing the action itself remains in the control of the subject, it is an action intentionally done for reasons, and to remain rational, the subject should act according to the result of his judgement. Thus, in cases of disagreement, the rational thing to do is not follow authority, follow one's own judgement instead.

This point – practical rationality requires congruity between one's judgement and one's intentions - is reinforced: instrumental rationality, always requiring that "one does what one has most reason to do," demands always balancing all the reasons there are for and against the action in question. This requirement is in tension with Raz's indirect instrumentalist strategy, requiring acting on protected reasons, not allowing such balancing. I have discussed in detail several critiques along these lines in the third part of my thesis, so I do not need to belabor the point further. Instrumental justifications meet serious problems in solving the "normative gap" problem. Thus justifications in terms of instrumental rationality do not easily fit authority with a practical exclusionary force

3.1.2. Beyond Mere Rationality: Substantive Shortcomings of the Instrumentalist Strategy

Apart from these general concerns with the rationality of acting on CiRs and ERs on the outcome-based instrumentalist model, there are substantive critiques, revolving around the issue of its adequacy as an account of the legitimacy of political authority. So, even if it turns out it is individually rational to act on the exclusionary directives of an instrumentally justified political authority, since it maximally improves one's conformity to one's reasons, this will not solve the legitimacy question. Meeting the rationality condition is not sufficient to show that one is under a duty to obey an otherwise rationally justified authority. The claim to legitimacy political authorities necessarily make – their claim to have a right to rule, correlates with a duty on the part of its subjects to indeed obey, if and when this authority is legitimate. Thus the instrumentalist strategy may succeed in establishing the rationality, but this does not thereby establish the moral

necessity of obeying such rationally beneficial authorities, and this is needed, if it is to be an adequate account of legitimacy.

A more concrete concern follows in the steps of the above one. Authorities meeting the instrumentalist legitimacy test may score rather poorly, when their justice is evaluated. This is the main fault with Raz's Service conception, according to Christiano:

> "The first problem with the normal justification thesis, primarily on its instrumentalist version, is that it *divorces* the issue of the justice of the authority from its legitimacy." Christiano (2004: 278, emphasis added).[318]

One need not stretch one's imagination too far to see how obeying grossly unjust authorities may still bring improved conformity to ones' reasons. The point here is not that rationality may legitimate what morality does not: it is not that one may be prudent to obey the unjust regime if threatened with harm. After all, NJT could be interpreted in thoroughly "moralized" fashion – authority is legitimate if and only if it brings improved conformity to one's moral reasons. Rather, the objection here is that authority may, even in bringing improved conformity to one's moral reasons, still be deeply unjust. Suppose an unjust state creates such conditions, in which the only possible way to follow one's moral reasons (avoid harming others, or try diminish the bad effects of its evil policies) is to cooperate with it by obeying its directives.[319] Though the morally best thing to do may indeed be obedience here (and this is a substantive, and controversial position), one would hardly want to grant legitimacy to such unjust authority.

I think a response to this charge may be available to the defenders of the Razian account (it needs further elaboration, though, to be conclusively evaluated). It could go along the following lines. Justice is relevant for determining the legitimacy of an authority even on this account, since subjects have strong reasons of justice. Through bringing improved

[318] The other one is that NJT ignores the moral significance of disagreement. Christiano (2004: 278, fn. 14) has not yet elaborated on his position that a non-instrumentalist interpretation of NJT is possible, may be even plausible. Raz's formulation of NJT permits, as I have shown in chapter 2 of my first part of this thesis, many different interpretations, but it is very difficult to see how it could be interpreted non-instrumentally. The driving idea of Raz's project is precisely to turn the authority-master into an authority-servant, making grand claims, but if and when legitimate, being just a good, useful instrument in the service of the well-being of its subjects.

[319] Such conditions often provide good novelists with the plot to develop their stories.

conformity to those reasons, authority scores points on this dimension as well.[320] I have touched on the issue of the relation of justice to Raz's account of authority in my chapter on normativity and coercion.[321] My claim there was that justice may influence the legitimacy of authority, but only indirectly - through the reasons of its subjects. It is (1) to the extent they have such reasons, (2) to the extent authority brings improved conformity to them, and (3) to the extent that bringing improved conformity to those and not to other reasons is more important, that there is a connection between the legitimacy of the authority and its justice. Though this situation should not be put in as dramatic terms as a "divorce" between justice and legitimacy, it should be recognised that the relationship is not sufficiently intimate either.

What a defender of the instrumentalist view has to further provide, is an account of the place of justice on this account of legitimacy. Maybe concerns with justice have lexical priority among the reasons of subjects and foregoing bringing improved conformity to them cannot be outweighed by bringing any measure of improved conformity to any other reasons? Or maybe reasons of justice do not allow to be served by bringing improved conformity to them - they do not allow "to be promoted" but rather require only "to be respected" (to use Scanlon's terms)? Or maybe justice is an altogether external condition to the legitimacy test, along the same logic as that of the autonomy condition, and does not work through the reasons of the subjects? These are important, difficult questions that need to be addressed. It is difficult to say in advance how successful the responses provided will be in meeting the above challenge. I suspect, though, that Raz's critics may well have slipped on this point: it may be possible to accommodate substantive justice on his model, if only indirectly – through improving conformity to subjects' justice-related reasons.

---

[320] Indeed, on the instrumentalist account, following Arneson (2003:123-124) one may distinguish between correctness standard and best result standards of evaluation. The first looks only at the outcomes narrowly taken, the other widens the perspective to include not only directly producing good, just decisions but even indirectly producing morally best results all in all. Thus a decision may not score as high on a certain dimension of value as some other, but still be preferable, since it may be conducive to bringing better moral outcomes overall - by instilling virtues in the citizens, prompting them act in morally better ways, etc. Raz's legitimacy conception could be understood on the second, wider understanding of outcome: producing best moral results overall, and in the long run. Notice that even on the narrower understanding, Raz's conception would not be vulnerable to the critique levelled by Christiano.

[321] See my chapter 3, section 3.2.2.1.

However, let me point out, that even if successful in this regard, the instrumentalist account of authority would still be employing an outcome-based test of legitimacy.[322] It is this aspect - its focus on outcomes, to the neglect of the procedures through which these outcomes are brought about, and not its complicated relation with substantive justice, which worries the theorists, searching for a plausible account of the authority of democratic institutions.[323] On such an account, the legitimacy of the democratic decisions, institutions, authority generally, crucially depends on employing democratic procedures for reaching binding decisions, rather than condition the validity of these decisions on necessarily reaching best outcomes. The intrinsic value of democracy is what validates the results from the democratic, collective decision-making process, and not their instrumental value - in bringing good, just, efficient outcomes.[324]

> "This instrumentalist conception ignores the intrinsic value of democracy. The legitimacy of rule is generally not judged exclusively, or may be even primarily, by its output, but rather by its input, that is by whether the regime has been determined, and is supported by the populace." Shapiro (2002a: 434)

One should be careful not to overstate the above points: it is unlikely that a plausible legitimacy test for democratic authority would exclusively rely on the procedural dimension, focusing on the inherent value of democracy alone, at the expense of forgetting any outcome considerations. Nevertheless, concern with procedures seems a necessary condition for having a plausible legitimacy test for democratic authorities: neglect of this concern is the main flow with the exclusively instrumentalist tests.

3.2. The Proceduralist Strategy: Valid CiRs by Democratic Pedigree?

[322] Recall the discussion in Part one of this thesis concerning the exclusively substantivist interpretation of Raz's NJT legitimacy test.

[323] Arguably, the point is more general. Not only defenders of democratic authority, students of any type of political authority do have a distinct interest in procedures. The procedural dimension thus is a relevant dimension for evaluation of political authority in general, not only that of democratic political regimes. Hershovitz (2003: 218), for example, claims (though he recognises he does not provide the necessary arguments) that Raz's NJT is inadequate as a legitimacy test for political authorities precisely on this ground – it is insensitive to this important procedural dimension.

[324] Certainly, this is not the only position a defender of democracy could take. Arneson (1993, 2003, 2004), to name but one example, defends an uncompromisingly instrumentalist account of democracy's value.

What I do in the rest of this chapter, is evaluate the success of an alternative, proceduralist strategy in explaining how one could have valid CiRs. If it succeeds, it will have a lot to recommend it. Its first advantage over Raz's outcome-based strategy is that it solves the normative gap/the break in transitivity of justification problems, without facing the difficulties Raz's own ERs solution suffered within a generally instrumentalist framework.[325] Thus it will provide support for a democratic type of authority on the ground that it is particularly well positioned in comparison to other types of authorities in that it can issue directives, giving valid CiRs to its subjects. Further, being a proceduralist strategy, it would give a solace to the democratic theorists, concerned with the disregard of procedures by the instrumentalist accounts of legitimacy. Authority's right to rule, on a proceduralist strategy, may be conditional on subjects' having a say, channeled through legitimating procedures, on whether the authority has this right. Thus such strategy may provide the framework for an account, however cursory, of the authority of democratic institutions, if these institutions embody the types of legitimating procedures, required by this strategy's test of legitimacy.

This strategy suggests that we leave aside the instrumentalist framework of justification - in terms of producing more value, or maximising good outcomes. The suggestion here is that in order to have valid CiRs, it is necessary to sever the connection of those reasons to *any outcome-based merits* (both of the directly commanded action itself and overall). This would require taking as relevant for determining the validity of CiRs only procedural merits.

3.2.1. A Case of CiRs?

To meet requirement (B) for content-independence – "no dependence on the evaluative properties of the content of the commanded action", it must be shown that the dependence of the validity of CiR on some procedural merits of acting as commanded,

[325] Most of the problems with ERs's validity I have discussed were traceable to the instrumentalist justification of authority. May be valid ERs are possible when other justifications for authority are advanced. Raz's account of the concept of authority (in which ER is of central importance) is independent from his own instrumentalist NJT and his conception of legitimacy generally. The coherence of the former should not depend on a "partisan" conception - this was Raz's (1989) response to Regan's (1989) complaint that Raz's account is not consequentialist enough. It still remains to be shown, if ER concept is coherent, that ERs can be valid as well within some non-instrumentalist framework. This is a task for further research.

does not involve the dependence of CiR on some merits of the content of this action. This requirement is met, if the authoritative directive is produced by a procedure with merits, where the merit of the procedure does not have an effect on - does not get "transmitted" to - the content of the action commanded.

1. Take an authoritative directive, produced by a democratic procedure.
2. Assume there is some merit in *producing* authoritative directives in this manner.
*3. (from 1. and 2.) This merit is procedural.
4. Independence of merit of content from merit of procedure: The *content* of an action required by an authoritative directive with democratic pedigree is not more "meritorious" than the content of an identical action, required by an alternative, non-democratic authoritative directive, even if the source of the former has merits (*ex hypothesi*) that the latter lacks. Since the contents by construction are identical, so are the merits of their contents as well.
*5.Thus the merit of the procedure is not transmitted to the merit of the content of the action required by its results.

This conclusion shows that requirement B for CiR is met: the validity of the authoritative reason is not dependent on the evaluative properties (the merit) of the content of the action, for which it is a reason.

This analysis shows that we do have a break of transitivity of justification in the case of authoritative directives by democratic (in procedural terms) pedigree: it testifies to the presence of CiRs. The reason for the adoption of the democratic procedure – that it uniquely realizes the value of justice, for example, though *validating* its result (the concrete authoritative directive) is not at the same time a reason for the concrete action required by this directive. The reason for the concrete action is that the authoritative directive requires it. Counterfactually, had the democratic decision-making procedure yielded a different authoritative directive, a different action would have been validly required. I conclude that if we have valid reasons for action, provided by authoritative

commands with democratic pedigree, they would indeed meet the requirements for content-independence and be CiRs.

3.2.2. Are CiRs with Democratic Pedigree Valid?

The interesting question, as before, is how acting on such CiR could be rational, since nothing in that reason points to the value of the action it requires. The action required might as well lack any value, even have negative value or be wrong. In this line of thoughts, one could ask whether indeed the "majority has the right to be wrong" [326] (if we assume that the legitimate democratic procedure requires majority decision-making rule), and if the answer is positive, how could one be justified to act on such majority's directives?

The above judgements "the required action lacks value" and "it is wrong, since the balance of reasons is against it" are grounded in *outcome-based* evaluations. Such judgements neglect the possibility that an action commanded by an authority with the right pedigree may have its value in its source, and not in the content of its outcome. If there is a value in performing actions, required as a result of *intrinsically valuable decision-making procedures,* it could be argued that an action, loosing along the dimension of outcome-values, may still have value and win out along the dimension of being required as a result of such procedures. In this limited sense, the majority (if majority rule is the valid procedure, required by the value of justice) may as well have "the right to be wrong," and one might still be justified to follow its directives.

All this seems straightforward enough.

It might be challenged, though, on several grounds. One is to deny that it makes sense to attribute value to an action on anything else but the value of its outcome.

As stated, this challenge does not seem to hold: identical in their outcome-value actions can have different value. Identical charitable acts (with identical outcome-value) can differ in their value depending on the reasons for which they were done - e.g. whether they were done to gratify oneself, or out of concern for the needy. Or whether they were

[326] Heidi Hurd's (1999: 101) answer to the question whether "the majority has the right to be wrong" is negative. She asks this question in the context of discussing precisely the issue whether content-independence can be made sense of if such a Ci directive was issued as a result of democratic procedure. She argues that CiR cannot be valid – neither on procedural, much less on substantive grounds. My discussion in the text was developed as a response to and a critique of her position.

done by following an egalitarian procedure for selecting the beneficiaries, or using some other, more sectarian criterion for selection – members of a particular church only, the group of anarchists only, etc.

However, the challenge is pressed further, even if it is agreed that there can be other than outcome-value, it might still be objected that ultimately ("at the end of the chain of asking why is this value a value") the procedural value is accounted for in terms of outcome-value. Thus conflicts between outcome- and other-value will ultimately be solved, by looking at the respective value produced in terms of better outcomes overall. That is, the procedural value gets sucked in the "consequentialist vacuum cleaner."[327] I will not try to respond to these claims: it will lead me too far (into the morass of the debates between consequentialist and deontological moral theories) from my concerns in this text. I will only try to provide some reasons why these objections need not be fatal for my argument for valid CiRs by democratic pedigree.

3.2.2.1. Democratic Proceduralism

At this point, one needs to distinguish between two positions, broadly described as proceduralist ones. The first takes what may seem the straightforward understanding of proceduralism – pure proceduralism "all the way down." On this view, democracy's value is entirely accounted for in pure procedural terms: as the realization of the value of pure procedural justice. This position seems the only viable alternative to the outcome-value account of democratic legitimacy, the only one that does not allow procedural value to get sucked into the consequentialist vacuum cleaner. Its success in this regard, however, comes at the expense of it being a plausible account of what we value in democracy: there is little to recommend this position.[328] I will leave it aside, and try to show there is a plausible alternative account of proceduralism, avoiding the shortcomings of this position.

---

[327] McNaughton and Rawling (1991). Arneson (1993, 2003) defends such a deflationary view: any assessment of procedure's fairness, even that of fair gambles, is driven by the likely or certain consequences of adopting this procedure.

[328] "We value political institutions because they make justice in society possible, because they advance the common good … Pure proceduralism is completely false to the practice of democratic citizenship." Christiano (2004: 269).

The second does not take proceduralism all the way down, but takes it as strictly required by some substantive value with fundamental importance. This is a much more plausible position. Thus the major reason to look at democratic procedures as potential sources of valid CiRs, on this conception, is that there might be a strong, substantive justification why the results from the democratic procedures *should not be properly* evaluated in terms of their outcome-value (or not exclusively in those terms). The claim is not the implausible one that there is no right and wrong (in outcome terms) apart from the procedure. It is, rather, that these often might be *irrelevant*. The type of justification needed for why is the outcome-value not a relevant dimension for evaluating the outcomes of democratic procedures, will point to the *intrinsic* value of democracy. The democratic decision-procedures (explaining their properties to yield right outcomes irrespective of their substantive merits) may be such that they uniquely realize certain intrinsically valuable properties – say, these of best expressing the principle of equal advancement/consideration of interests.[329]

Doubts on the plausibility of this latter type of democratic proceduralism may have to do with concerns that it lacks the potential to resist the consequentialist pull. The success of such theory, then, depends on persuasively showing that the democratic procedure is a unique realization of a fundamental value, which cannot be compromised or substituted for by scoring high on an alternative value.[330] Showing this is a rewarding theoretical exercise. If there is a successful argument along these lines, it will establish that democratic procedures uniquely realize an inherent value with fundamental importance. If these procedures indeed are so justified, then they will have certain properties, in virtue

[329] There is a considerable disagreement among the theorists of democracy precisely what kind of decision-making procedure do the principles of equal treatment, consideration, advancement of interests,… favour. There is almost unanimous agreement that these principles provide an argument for political equality (as a minimum requiring one person - one vote), and that they ground a right to participation in democratic decision-making, both at its deliberation and its decision stages (Arneson takes an exception concerning the right to democratic "say"). It is a hotly contested issue, however, whether majority rule (equal weighing of votes) is the unique realization of political equality. Equal weighing of votes is not desirable in all circumstances, in the same way as the anonymity and the neutrality requirements are not absolute desiderata for having fair decision-making rules: majority rule is just a special case and not a uniquely fair procedure. Janos Kis (2003: 65-74) persuasively argues for the symbolic significance (because expressing equal concern and respect for each person) of one person-one vote, which does not, however, extend to equality of vote, and thus does not necessarily favour majority rule in deciding all issues.

[330] This seems to be one way Arneson (2003) himself urges the defenders of the intrinsic value view of democracy to proceed, in order to present a real alternative to the consequentialist, broadly outcome-based view he defends.

of which they are such a unique realization of this value. It is precisely these properties, responsible for their being the unique realization of an intrinsic value, which grant legitimacy to the results of democratic procedures.

When seen in this perspective, this renders misleading the granting to majority of a "right to be wrong." Whether the decision majority has reached is wrong or not, on such an account, should be established by considering whether this decision has the right pedigree. This rightness of the decision is due to the fact that the procedure through which it was produced, realizes the intrinsic value of democracy (itself a unique realization of a fundamental value). The decision is rendered valid by the inherent properties of the procedure, realizing this value.

Thus establishing the right pedigree might be all one needs to do to establish that the decision in question is valid, that the action it requires has a value. One is thus rationally justified in acting on this decision: no other value except that procedural one, plays out at this stage. Put in somewhat confusing terms,[331] the pedigree at this stage is partly constitutive of the value of the decision.

What justification could be offered for such a position? What value(s) could render democracy, with its procedural component, intrinsically valuable? This is not the place to address these complex and important questions. What I instead do in the rest of this section, is indicate what I think is the type of argument that needs to be made to make a plausible case for such a position.

A plausible argument for an irreducibly proceduralist component in justified government, could be provided by a substantive egalitarian theory of social justice,[332] which proceeds along the following lines. Start with the fundamental value of equal advancement of interests: this is the basic principle of social justice. Under certain conditions (of diversity, cognitive bias and disagreement, for example) it requires that the equal

[331] The position discussed here is different from and should not be confused with a "constructivist" account of democracy, relying on unanimity decision-making rule, and only resorting to majority rule as a second-best solution. Apart from being practically unattainable, unanimity decision rule is undesirable since it violates political equality in a political world (such as ours) which has not evolved cooperatively from a just initial position, it is a mistake to associate it with an egalitarian position of the sort discussed in the text.

[332] This account is offered in Christiano (2004). It is a reformulation and further elaboration of his account, earlier defended along somewhat different lines in Christiano (1996).

advancement of interests be publicly realized. This means that the judgement of each is to be respected and the principle of weak publicity - realized: each should be *able to* see that he is treated in accordance with the principle of equal advancement of interests. Democratic decision-making on issues where interests conflict, and thus reaching decisions is morally required, is the uniquely public way to realize justice – the uniquely public way for equally advancing the interests of all. Democracy is of intrinsic value, then, because it is the unique way of resolving conflicts of interests and substantive disagreements, while remaining faithful to the fundamental principle of public equality.

This position is far from being uncontroversial. One could argue against it at each step of the argument. Why publicity is important? It is not an independent value, but, rather it is entirely parasitic on the value of justice: why the fuss? Why judgement is taken as a reliable proxy of interests? Is it not the case one has an important interest in society reaching the best judgement on how the conflict of interests is resolved, and not that one's own judgement necessarily determines this collective judgement, if this increases the likelihood it will be wrong? Why is disagreement taken to be of such fundamental importance – does not its presence simply show that some are wrong, and there is no value in taking their judgements seriously? Respecting persons need (may) not involve respect for their mistaken judgements[333]….These are some of the most pressing questions to the solutions offered. Maybe other, better arguments could be supplied in their place. Or maybe this solution proves altogether not successful.

Evaluating the success of this theory as an account of democratic authority (it is advanced as precisely such an account)[334] is certainly a highly rewarding task. This is not my task here, however. This theory is just a placeholder here. Any adequate theory with the same structure: fundamental value, which upholds the intrinsic value of democratic procedures as a unique realization of that fundamental value, would do. All I needed was to show that there is at least an initially plausible theory, on which democratic procedures have irreducible, intrinsic value as the unique realization of a value with fundamental importance. The value of equal advancement of interests certainly is such a value.

[333] Raz (1998), Arneson (2003).

[334] The title of Christiano's (2004) article is "The Authority of Democracy," and it is indeed the questions about the authority of democracy, that are being discussed there.

The obvious problem is that the argument from democratic proceduralism seems to rely on the implausible claim that there are some issues, on which there *are no* independent standards for evaluating the correctness of the decisions. This is so on a pure procedural justice understanding of proceduralism, or the all-the-way-down proceduralism I dismissed as an unattractive account of the value of democratic procedures. Is the moderate democratic proceduralism safe from this objection?

I think it is. The claim here is not that there are no standards outside the procedure, against which to judge its results. Rather, the claim is that there are such standards, but they are irrelevant at this stage: the fundamental value, whose unique realization is embodied in the democratic procedure, protects the procedural value, excluding the role other considerations can play at this stage. Independent evaluation exists, but it is not appropriate here.

Again, this solution is far from being unproblematic. What could be the justification for holding some value to be of such absolute importance, so that not to permit independent evaluation to apply to the results of the democratic process? It is immensely implausible, that concerning all issues, such independent standard of evaluation will be properly irrelevant.

The democratic proceduralist position, then, will need some relaxing, so that different spheres are delineated, in which different standards of evaluation play out. To preserve the coherence of the position, it could be further argued, that restricting the domain where the procedure is all that counts, may be what is required by the imperative of properly respecting the same fundamental value, which upholds the value of democratic procedure itself. Again, there are bound to be objections to this line of argument.[335] There are many unanswered questions as well. Does the fundamental value give a determinate answer on which issues of collective concern fall into which sphere, and what is their relative weight in evaluating the overall merit of government (both procedurally and substantively)? And even if the fundamental value does give a determinate, principled answer, what does the fact of disagreement precisely on that issue - *whether* there is such

[335] This position is again advocated by Christiano (2004), who seems to be addressing similar charges against his democracy argument, developed earlier in Christiano (1996) as being overly proceduralistic. It is difficult to evaluate the success of the argument on the basis of a programmatic article. Further discussion should await the publication of his monograph (Christiano 2005)

an answer and *what* is it - say about whether a decision on it as well should be left to the public democratically to settle?[336] Would not the latter solution of leaving it for the public to democratically decide, lead to contingent, arbitrary results, constantly shifting the boundaries between the spheres, etc…, etc..?

All I can say at this point, is that the position discussed is not obviously implausible. There seem to be nothing obviously implausible to say that within a certain sphere the only relevant values may be the procedural ones, if there is a strong justification for this being the case.

4. NJT's Filtering Role: The Full Legitimacy Test and Liberal Democracy

This provisional result helps me venture an answer to the puzzle that prompted this long discussion: how could one have a valid reason to follow the results of a democratic procedure, irrespective of their merits; how could such results be binding? I have argued that, as stated, these questions are misleading, since they neglect the possibility of having a decision, whose merit is not to be determined without taking into account the process, by which it was reached. If this process is a unique embodiment of an inherent value with a considerable importance, then it will be rational (may be even obligatory – if the inherent value at stake is a source of moral requirements) to accept it as justified and act on the decision, yielded by this procedure. The fundamental value of equal advancement of interests, for example, may be such a source of moral requirements: then it may be indeed morally required and not simply rationally advisable to act on the results of the procedure, its value strictly required.

This answer also helps us see how democratic procedures could yield valid CiRs: such reasons that seemed riddled with difficult to track and resolve puzzles. As I have attempted to show, these reasons are particularly puzzling on the instrumentalist type of justification for their validity. Such justification conditions the validity of those reasons on producing maximally valuable outcomes. But if producing maximally good outcomes is the justification, how could it require acting on reasons irrespective of their bringing about such maximally good outcomes? The ERs solution, that is meant to help resolve

[336] This is famously argued for by Waldron (1991, 1998, 2001).

this puzzle, does not succeed in the case of instrumental justifications either. It could plausibly be argued (as I did in this chapter and in Part Three of this thesis) that if the justification for taking those reasons as valid is again instrumentalist (based on maximising instrumental rationality), it does not establish it is rationally justified to act on them. NJT and the whole instrumentalist Service conception of legitimacy, it seems, should be abandoned as wholly misguided.

I have also tried to show that when the instrumentalist framework is left behind, it is possible to show that it is justified, may be even morally required to treat the authoritative directives, resulting from a democratic procedure, as valid CiRs. Further, it could be argued that these authoritative directives can be treated as exclusionary reasons for action. Disregarding one's own reasons may be morally required, if one has a moral obligation to do so, deriving from the fundamental substantive value that is being uniquely realized through the democratic procedures for reaching collective decisions.

And here comes the problem: this neat, beautiful even, solution cannot be correct. The picture of a thoroughly non-instrumentally justified political authority, providing its subjects with valid protected reasons for action, which owe their validity to nothing else but procedural merits, is not true to our considered convictions about what is a legitimate institutional arrangement. An "ideal world" democratic society with all authoritative decisions having the requisite democratic pedigree (and as such all being valid), thus having procedural and no substantive merits whatsoever (or, worse, having just substantive de-merits) if it worked at all, would be a very bad polity indeed - a utopianist's nightmare. Delivering substantively just, good, efficient outcomes, is necessarily part of the legitimacy story, we justly believe.

This suggests that an instrumentalist - NJT legitimacy test should not be entirely abandoned. It should, however, be thoroughly reinterpreted (dropping the maximising interpretation of instrumental rationality in favour of satisficing one, for example) and downgraded into playing the type of "filtering role" I briefly discussed in the second chapter of this thesis. It could provide the test of substantive adequacy any candidate for a plausible conception of legitimacy should meet. NJT thus will be a minimally necessary condition for legitimacy. There is, however, a need for a further, non-instrumental, non-

outcome-based type of legitimacy test: the instrumental component cannot by itself legitimate authority – authority would not yield valid protected reasons for action. It is the conjunction of the instrumentalist minimally necessary condition, with this second component, which is a must for an adequate conception of legitimacy, on which authority could be a source of such valid reasons. This means that one needs to opt for a *dualistic conception* of legitimacy, and face all the problems with it.

Thus the drift of my whole argument, let me stress it here, points to the conclusion that NJT cannot furnish a fully sufficient test of legitimacy. It is not that other considerations – such as the autonomy condition, may occasionally deflect from the legitimacy, when satisfied, it otherwise confers on a government. Even absent countervailing considerations, it would not be such a sufficient test. This means NJT is not a threshold test: even other things being equal, it is not sufficient. It is a filter: any adequate conception must meet it, but in addition, something different altogether is needed for legitimacy.

The problem, I have argued, is that meeting the requirements of an instrumentalist legitimacy test does not meet the adequacy requirements of Raz's conception of practical authority - an authority with practical exclusionary force. The adequacy requirement is that legitimate practical authority is an authority that provides valid protected (CiR and ER) reasons to its subjects. Any successful legitimacy conception should have the capacity to yield justifications, which could meet this adequacy requirement. The instrumentalist justifications, Raz's NJT included, fail to support the validity of such reasons, thus failing the adequacy test: they do not justify authorities with practical exclusionary force. Legitimate authority of this type is impossible on purely instrumental grounds. Hence NJT as a sufficient test of legitimacy is in principle inadequate.

This is a comforting conclusion: that there is nothing in Raz's instrumentalist conception of legitimacy that would favour the type of political arrangement we consider morally best– the liberal-democratic one, was a constant source of cognitive discomfort. On NJT, an ideal enlightened autocracy, producing substantively maximally good outcomes, could be more legitimate, than a liberal-democratic government that, though generally producing similar outcomes, occasionally performed sub-optimally, due to the presence

of democratic rules for collective decision-making. There was nothing in NJT that could show why these alleged "shortfalls" of the democratic rule, are not necessarily shortfalls. As it turns out, we should not have been really too much concerned, since NJT is anyway inadequate as a sufficient, full test of legitimacy.

But the result I have reached is far from being entirely satisfactory. What has to be shown, first, is that NJT as a minimum necessary condition for an adequate legitimacy conception is *compatible* with an inherent-value justification, on which a procedural (not outcome-based) realization of this value is required. I have claimed that specifically such type of justification could render practical authorities legitimate and the protected reasons they yield – valid. Maybe NJT is incompatible with such justifications? Nothing said here points in this direction. Establishing that a unified account of liberal-democratic authority, which relies on a dualist (both instrumental and in terms of inherent value) type of justification for political authority is coherent, is the next topic that need be addressed. I expect that such a dualistic account will meet serious problems: there will be a constant pull toward one or the other value, since the instrumental and the inherent value-components could be expected to constantly come into conflict. But for now I leave it at that.

Secondly, for us to rest content that our preferred liberal-democratic political order is a political arrangement, on which authority can be exercised *fully legitimately*, it needs to be shown that this type of political order realizes in its institutions, decision-making procedures, etc., the type of inherent value I claimed is necessary for full legitimacy. The type of argument needed for this is a substantive one: though it will have to be congruent with the formal arguments advanced in this chapter. This argument will not automatically favour this type of political arrangements over others, however: other political arrangements may also uniquely realize an inherent value of the required above type (though I doubt an enlightened autocracy will qualify as fully legitimate).

Thus, thirdly, a defense for a liberal-democratic political order - the one we believe best, morally speaking, from the available alternatives - should establish that the inherent value necessary for conferring full legitimacy, not only is uniquely realized in this type of institutional arrangement, but it is the only value of the required type, or is the only value

that could be realized. Again, the arguments needed will be substantive, and certainly controversial.

These last two points show why my argument here cannot be conclusive: I have provided only the matrix for the arguments needed. As such, my argument is complete, since I was looking for the types of arguments, required to provide justification for a liberal-democratic authority. It would be conclusive, if the substantive arguments were also provided. Attempting to do that, however, should await another occasion.

5. Conclusion

In conclusion, let me briefly summarize what I have been up to in this chapter. Employing a plausible definition, with a sufficient requirement for CiR, I have addressed the main problem with the coherence of this concept – that of the "normative gap" (between what one ought to do, and the good of doing it). I have claimed the problems with it, though serious, are not unsurpassable, and need not detain us from our main interest: identifying conditions, which could establish the validity and the rationality of acting on such reasons. Two strategies (an instrumentalist, outcome-oriented and a proceduralist one) of identifying such conditions were outlined, and after indicating considerable difficulties with the first, notably - its failure to provide satisfactory solution to the "normative gap" problem - the second was elaborated in some detail and was found satisfactory in this regard. It was established that a plausible account of the validity and the rationality of acting on CiR could be offered if they are provided by a democratic authority, employing decision-making procedures which uniquely embody a certain inherent value of fundamental importance – that of equal advancement of interests, say. A brief discussion of this substantive conception – its advantages and shortfalls, was provided.

Next, the failure of the outcome-based instrumentalist strategy prompted discussion of the adequacy of Raz's Normal Justification Thesis as a sufficient legitimacy test. I have argued that it fails as such test since it fails to identify the conditions, under which political authority could be practically legitimate – could provide valid protected (CiR and ER) reasons to its subjects. It was claimed that nevertheless NJT is a minimally necessary condition for any conception of legitimacy. An adequate conception thus

would combine an instrumentalist and an inherent-value (strictly requiring proceduralist realization) components. At the end of the chapter, a brief list and a short discussion of the problems for further research were provided.

Conclusion

A liberal-democratic type of political order is commonly, if not universally, believed to provide the best institutional arrangement for contemporary societies. And though some political theorists argue that there is an irreducible conflict between the values of liberty and equality at its foundations, the common-sense notion seems to favour the opposite position - both values are important and do not conflict; they both can be satisfied without compromise. Our best philosophical theory seems to side with the common-sense notion: a liberal-democratic political arrangement harmoniously accommodates within a common institutional framework these fundamental values.

The best theory, however, does not give an immediate answer to one of the leading questions that prompted my long journey. It was what kind of argument is needed to show that the authority in such valuable political arrangement can be legitimate. The hypothesis that a reason-based justification for a liberal-democratic type of authority may be the adequate type of argument was thus in the focus of the discussion in my thesis. Specifically, I tested the capacity of the most sophisticated and fully developed contemporary conception of a reason-based justification for authority, to offer such a justification. This is the Service conception of legitimate authority, advanced by Joseph Raz .

The interpretive part of my thesis was devoted to discussing in considerable detail the main concept of Raz's account of practical authority (itself in the background of Raz's Service conception of legitimacy) , carrying the burden of explaining the binding force of practical authority's valid directives. This is the protected reason for action concept. The problems with the coherence of its two elements – the content-independent and the exclusionary reason, were discussed in detail. I have shown that the distinctness of CiRs is best drawn out if it is characterised by the "no dependence on evaluative properties of the action" requirement. I have also demonstrated how this characterisation better illustrates the "normative gap" problem with the coherence of this concept. I have recognized there are several as yet unanswered problems with the conceptual coherence of the ER component as well. I have also suggested there are some further, not yet fully appreciated problems with determining the weight of this type of reason. I have indicated that this latter problem may challenge one of the important features of this concept,

accounting both for its own plausibility as well as explaining how legitimate authorities are at all possible – that it has a limited scope of application.

I have then suggested that we turn away from these conceptual puzzles and move into the site, where the real action is: justification. Raz's essentially instrumentalist "Service conception" of legitimacy was introduced together with the main interpretations of its legitimacy test – the normal justification thesis (NJT). I have then argued that on this type of justification it is difficult, if not impossible, to pinpoint the way in which legitimate authority can make practical difference to how its subjects ought to act. I have also identified the main problem with this test: it cannot account for the sense in which practical authority, when legitimate, makes practical difference to what its subjects ought to do by providing them with a moral duty to obey it specifically. My claim was that authority may give rise to an instrumentally justified hypothetical rational requirement to obey, but it is far from clear in what way authority can, on NJT, turn this rational requirement into a *moral duty* to obey authority's directives.

I have moved from the site of "general justification" for authority onto the ground where my main interest lies – political authority. I have tested the congruity of some of the central features, according to Raz, of political and legal authority - the normative claim to legitimacy, the claim to supremacy over all other normative domain and their extensive use of coercion. Furthermore, I have tested whether they can be accommodated within Raz's Service conception of legitimacy. I concluded that the compatibility of the normative claim authority necessarily makes with the state's extensive use of coercion, has not been securely established. The disjunctive view, challenging their compatibility, has not been refuted. I have also claimed that this result may not be restricted, as it was meant, to the case of some non-instrumental reasons to obey, introduced to fill in the gaps in the intermittent picture of an instrumentally justified practical authority. Rather, I have shown it raises a more serious concern. If practical authority is described as a source of duties (even if not ultimately so) and its being such a source is indeed undermined by the state's use of coercion, this means that on this conception either the state could not impose duties of obedience, i.e. could not be practical authority, or it could not use coercion. If neither is acceptable, one either has to abandon the practical model of

authority, or has to provide successful arguments against the disjunctive view. I have claimed the latter has not yet been done.

I have also argued that another feature of political and legal authority – its claim to supremacy over all other normative domains - does conflict with the autonomy condition of the Service conception: that it is often more important to decide oneself rather than decide correctly. Of the two theses that could support the autonomy condition – Dworkin's famous endorsement constraint thesis and the agent-relative reasons thesis, I found the first seriously flawed, and offered a defense for the latter. I concluded that the plausibility of the autonomy condition as supported by the agent-relative reasons thesis provides a challenge to Raz's conception of political and legal authority and their justification in a liberal-democratic political order. My claim was, first, that the autonomy condition on the agent-relative reasons thesis interpretation, casts doubt on whether political authority can make a bona fide claim to supremacy over all other normative domains. Secondly, I argued that a *liberal-democratic* type of political authority specifically, on which the restrictions of the autonomy condition are taken even more seriously, is by virtue of that even less capable of making such a bona fide claim. Since making this claim is, on Raz's conception, an essential feature of law, and by implication, of state authority, this shows not only the tensions within Raz's own conception as an adequate conception of political authority in general, but indicates the problems with applying this conception to the case of liberal-democratic type of political authority specifically. It is plausible to assume that a central, defining feature of this type of political authority in particular is that it refrains from making such an overboard, comprehensive claim to supremacy over all other normative domains. Thus the strong conclusion reached was that making such a claim cannot be a central feature of the concept, since the case of the liberal-democratic type of political authority falls within the core of the concept of political authority. I have, thirdly, suggested that the internal coherence of Raz's conception of the legitimacy of political authority can also be questioned. This is so because the general drift of his Service conception of legitimacy – obedience to authority is justified when licensed by morality, or practical reason more generally, goes against this normative supremacy claim as a central feature of legal and political authority.

These strong critiques against Raz's model, I claimed, do not immediately warrant disqualifying it as an adequate conception of political authority. One of its major advantages is that it may resolve the puzzles of rationality, involved in our common-sense notion of practical authority.

Thus the third part of my thesis was devoted to this major issue: does Raz's Service conception have indeed this rationality advantage. The general answer, given on this conception by it test of legitimacy - NJT, establishes the compatibility of authority with rationality. Following instrumentally justified authority – one that brings improved conformity to one's own reasons, is rational. Thus the focus of the discussion was whether it is indeed individually rational to decide to follow an instrumentally justified authority, if to follow authority means to take its directives as protected reasons for action - as it is on Raz's model of practical authority. I have dismissed a negative answer to this question, drawing on an alleged analogy of the authority case with that of Gregory Kavka's Toxin Puzzle. Nevertheless, the analysis of this possible analogy helped identify a different though closely connected problem for the rationality of deciding to follow such type of authority. Rationality seems not only to permit, but even to require following authority (when the latter is instrumentally beneficial), but only in case room for exceptions is allowed. However, this breeds "instability:" in cases of disagreement with authority, one is always tempted to treat them as such exceptions, and as a result one may end up being worse off than if he never decided to follow authority in the first place. I have shown that neither Raz's own conception, nor the application to it of Michael Bratman's "modified sophistication" decision-making strategy offer solution to this problem. I have also demonstrated the inadequacy of Scott Shapiro's Constraint model of authority, as a response to these problems of rationally following an instrumentally justified authority.

These failures, when added to the previously raised general critique against Raz's instrumentalist justification of political authority - of not being true to our notion of authority as implying a *moral* duty to obey legitimate political authority, acting within the bounds of its jurisdiction - prompted attempting to go beyond a generally instrumental justification in the case of political authority. I tried to do this in the fourth, concluding part of my thesis.

Thus I next showed that one of the reasons it was particularly difficult to demonstrate the rationality of following authority, was that on instrumental, outcome-based grounds only, it is not rational to treat an authoritative directive as CiR. This is so, because on such grounds it is difficult to close the normative gap, opening with the presence of authoritative directives between the reason - one ought to obey "the say so of an authority," - and the good of acting as required by that reason. I have argued, instead, that CiRs can be valid, and acting on them – rational, on non-instrumental, non-rationality maximizing grounds. If, for example, such reasons are provided by a democratic authority, employing decision-making procedures, which uniquely embody a certain inherent value of fundamental importance, these reasons can be valid and acting on them – rational. Whether democratic authority does indeed provide such valid reasons, depends on the success of a substantive argument that democratic procedures are the unique realization of an inherent value of fundamental importance. Thus though my argument for the legitimacy of democratic authority is complete in itself - it is the type of authority that could in principle provide valid CiR to its subjects, - this argument is inconclusive. There might be no fundamental values that would strictly require to be uniquely realized in a democratic procedure. Or there might be more than one such fundamental value, where only some of them require realization in a democratic procedure, while others could be realized through other, not specifically democratic procedures.

Nevertheless, apart from the theoretical advantages - explaining away one of the rationality puzzles involved in our concept of authority, - my result imposes considerable limits on the types of political orders, which can be shown to enjoy legitimate authority and provide to its subjects valid reasons of the required type. If an otherwise beneficial political order fails to employ procedures, uniquely realizing some fundamental value, such regime would fail to be legitimate, because it would fail to provide such valid reasons to its subjects. Further, even if democratic political order fails to always bring its subjects maximally improved conformity to their own reasons, the fact that it gives them valid reasons of the required type may confer legitimacy on it. This clearly favours it over regimes, which are otherwise "more successful" in purely instrumental, maximising-rationality terms.

This result also showed why Raz's Service conception of legitimacy is in need of a revision. Its legitimacy test – NJT, is not a full, or a sufficient test of legitimacy. Legitimate authority is the authority that gives valid reasons of a special type - protected reasons with a CiR and an ER component. Since on NJT grounds it cannot be shown how authority could give such valid reasons, this establishes that NJT is inadequate as such a full legitimacy test. However, I suggested that NJT could furnish a filter test that any regime should be able to pass, in order to qualify as being legitimate on further grounds. Thus meeting Raz's NJT should be seen as a necessary but never sufficient condition for the legitimacy of the authority of a political order.

The direction for developing a full test for the legitimacy of a liberal-democratic authority I see along the following lines. First, the filter NJT test should be passed: some adequate level of this regime tending to bring improved conformity to subjects' own reasons is reached. Next, the requirements of the autonomy condition should also be met. It is the subjects themselves, who decide on a wide range of issues concerning only them, whether deciding correctly is more important than deciding for themselves. This means legitimate authority should provide them with opportunities to act directly on their agent-relative reasons for action as they see fit, by protecting their rights against others and against the state. This is the liberal component of the legitimacy test. And, finally, the legitimacy test is only complete when the results of a democratic, collective decision-making process provide subjects with valid reasons to submit to those results.

In conclusion, my claim is that this legitimacy test would still offer a reason-based justification for submitting to authority's demands for obedience. It is not subjects' agreement with the liberal-democratic authority's directives, or their agreement that the requirements of this legitimacy test are indeed met, that render such authority legitimate and require their obedience to its directives. It is rather their own reasons, which authority promotes, serves as a good protection to, or pays due respect to, that require following such an authority.

Bibliography

Alexander, Larry. (1990) "Law and Exclusionary Reasons" in *Philosophical Topics* 18: 5-22

Alexander, Larry. (1999) "Can Law Survive the Asymmetry of Authority?" in Linda Meyer (ed.) *Rules and Reasoning: Essays in Honour of Fred Schauer*, Hart Publishing, Oxford

Arneson, Richard. (1993) "Democratic Rights at National and Workplace Levels" in David Copp, Jean Hampton and John E. Roemer, (eds.) *The Idea of Democracy* Cambridge University Press, Cambridge

Arneson, Richard. (1999) "Human Flourishing versus Desire-Satisfaction," in *Social Philosophy and Policy*, v. 16 no1.

Arneson, Richard. (2002) "Review of Dworkin, Ronald. *Sovereign Virtue: The Theory and Practice of Equality*" in *Ethics* 111: 367-371

Arneson, Richard. (2003) "Defending the Purely Instrumental Account of Democratic Legitimacy" in *The Journal of Political Philosophy* 11: 122-132

Arneson, Richard. (2004) "Democracy Is Not Intrinsically Just" in Keith Dowding, Robert Gooding and Carol Pateman, (eds.) *Justice and Democracy*, Cambridge University Press, Cambridge

Bratman, Michael. (1987) *Intention, Plans and Practical Reason* Harvard University Press Cambridge MA

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 University Press, Cambridge

Bratman, Michael. (1999) *Faces of Intention: Selected Essays on Intention and Agency*, New York

Broome, John. (2000) "Normative requirements" in *Normativity*, edited by Jonathan Dancy, Blackwells, Oxford

Broome, John. (2004) "Reasons" in *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, edited by Jay Wallace, Michael Smith, Samuel Scheffler and Philip Pettit, Oxford University Press, Oxford

Besson, Samantha (2005, forthcoming) *The Morality of Conflict. Reasonable Disagreement and the Law*, Hart Publishing, Oxford

Buchanan, Allen. (2002) "Political Legitimacy and Democracy" in *Ethics* 112: 689-719

Chang, Ruth. (2004) "Can Desires Provide Reasons for Action" in *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, edited by Jay Wallace, Michael Smith, Samuel Scheffler and Philip Pettit, Oxford University Press, Oxford

Christiano, Thomas. (1993) "Social Choice and Democracy" in *The Idea of Democracy*, (ed. by Copp, Roemer and Hampton), Cambridge University Press, New York

Christiano, Thomas. (1996) *The Rule of the Many: Fundamental Issues in Democratic Theory*, Westview Press

Christiano, Thomas. (2004) "The Authority of Democracy" in *The Journal of Political Philosophy* 12, no 3: 266 - 290

Christiano, Thomas. (2005, forthcoming) *Democratic Equality* , Oxford University Press, Oxford

Clayton, Matthew. (2002) "Liberal Equality and Ethics" in *Ethics* 113: 8-22

Coleman, Jules. (2001) *The Practice of Principle* Oxford University Press, Oxford

Cunliffe, John and Andrew Reeve. (1999) "Dialogic Authority" *in Oxford Journal of Legal Studies* 19: 453 - 465

Dan-Cohen, Meir. (1994) "In Defense of Defiance", in *Philosophy and Public Affairs* 23, no 1: 24-51

Dancy, Jonathan. (1993), *Moral Reasons*, Blackwell, Oxford

Dancy, Jonathan. (2000) *Practical Reality*, Oxford University Press, Oxford

Darwall, Stephen. (2002) *Welfare and Rational Care*, Princeton University Press, Princeton

Durning, Patrick. (2003) "Joseph Raz and the Instrumental Justification of a Duty to Obey the Law" in *Law and Philosophy* 22: 597 - 620

Dworkin, Ronald. (1987) "What is Equality?" Part 4 "Political Equality" *University of San Francisco Law Review* 22

Dworkin, Ronald. (1989a) "Foundations of Liberal Equality", *Tanner Lectures on Human Values X*I. Salt Lake City

Dworkin, Ronald. (1989b) "Liberal Community", *California Law Review* 77

Dworkin, Ronald. (2000) *Sovereign Virtue: The Theory and Practice of Equality*, Harvard University Press, Cambridge, Mass.

Edmundson, William A. (1993) "Rethinking Exclusionary Reasons," a review of Raz's Postscript to the second edition of *Practical Reason and Norms*, in *Law and Philosophy* 12: 329-43.

Edmundson, William A. (1998) *Three Anarchical Fallacies: An Essay on Political Authority*, Cambridge University Press, Cambridge

Edmundson, William A. (2002) “Social Meaning, Compliance Condition, and Law’s Claim to Authority” in *Canadian Journal of Law and Jurisprudence,* XV, no 1: 51-67

Edmundson, William A. (2003) “Locke and Load” in *Law and Philosophy* 22: 195-216

Edmundson, William A. (2004) “State of the Art: The Duty to Obey the Law” in *Legal Theory* 10: 215-259

Elster, Jon. (1984) *Ulysses and the Sirens: Studies in Rationality and Irrationality*, Cambridge University Press, Cambridge

Elster, Jon. (2000) *Ulysses Unbound*, Cambridge University Press, Cambridge

Finnis, John. (1980) *Natural Law and Natural Rights*, Clarendon Press, Oxford

Flathman, Richard E. (1980) *The Practice of Political Authority: Authority and the Authoritative*. The University of Chicago Press, Chicago

Galston, William. (2001) “The Obligations of Equality - Review of Ronald Dworkin's *Sovereign Virtue: The Theory and Practice of Equality*” in *Review of Politics* 63: 607-611

Gans, Chaim. (1986) ”Mandatory Rules and Exclusionary Reasons,” in *Philosophia* 15: 373-94

Gans, Chaim. (1992) *Philosophical Anarchism and Political Disobedience* Cambridge, Cambridge University Press, Cambridge

Gardner, John. (2001) “Legal Positivism: 5 1/2 Myths” in *American Journal of Jurisprudence* 46: 199 - 229

Gauthier, David. (1996) “Commitment and Choice: An Essay on the Rationality of Plans” in *Ethics, Rationality, and Economic Behaviour,* ed. by Francesco Farina, Frank Hahn, and Stefano Vannucci, Oxford University Press, Oxford

Gauthier, David. (1998a) “Intention and Deliberation” in Peter Danielson (ed.) *Modeling Rational and Moral Agents,* Oxford University Press, Oxford

Gauthier, David. (1998b) “Rethinking the Toxin Puzzle” in *Rational Commitment and Social Justice. Essays for Gregory Kavka,* ed. by Jules Coleman and Christopher Morris, Cambridge University Press, Cambridge

Green, Leslie. (1989) “Law, Legitimacy and Consent” in *Southern California Law Review* vol. 62: 795-825

Hardin, Russell. (1999) “Do We Need Trust in Government?” in Mark E. Warren (ed.) *Democracy and Trust*, Cambridge University Press, Cambridge

Harman, Gilbert. (1975) “Reasons,” Critica 7 (1975): 3-13; reprinted in Joseph Raz (ed.), *Practical Reasoning* , Oxford University Press, Oxford

Harman, Gilbert (1998) “The Toxin Puzzle” in Jules Coleman and Christopher Morris (eds.) *Rational Commitment and Social Justice. Essays for Gregory Kavka*, Cambridge University Press, Cambridge

Harre, Rom. (1999) “Trust and its Surrogates” in Mark E. Warren (ed.) *Democracy and Trust*, Cambridge University Press, Cambridge

Hart, H.L.A. (1961) *The Concept of Law*, Oxford University Press, Oxford

Hart, H.L.A. (1982) “Commands and Authoritative Legal Reasons” in Hart, H.L.A., *Essays on Bentham*, Clarendon Press, Oxford.

Hershovitz, Scott. (2003) “Democracy, Legitimacy and Razian Authority” in *Legal Theory* 9: 201-220

Himma, Kenneth Einar. (2000) “H.L.A. Hart and the Practical Difference Thesis” *Legal Theory* 6: 1 - 43

Himma, Kenneth Einar. (2001). “Law's Claim of Legitimate Authority,” in Jules L. Coleman (ed.), *Hart's Postscript: Essays on the Postscript to the Concept of Law*. Oxford: Oxford University Press, pp. 271-309.

Hurd, Heidi. (1999) *Moral Combat* Cambridge University Press, Cambridge

Hyland, James. (1995) *Democratic Theory: The Philosophical Foundations*, Manchester: Manchester University Press

Kagan, Shelly. (1989) *The Limits of Morality* Clarendon Press, Oxford

Kavka, Gregory. (1983) “The Toxin Puzzle” in *Analysis* 43: 33-6.

Kimel, Dori. (2003) *From Promise to Contract: Towards a Liberal Theory of Contract*, Hart Publishing, Oxford and Portland, Oregon

Kis, Janos. (2003) *Constitutional Democracy*, Central European University Press, Budapest, New York

Kramer, Mathew. (1999) “Requirements, Reasons and Raz: Legal Positivism and Legal Duties” in *Ethics* 109: 375-407

Kramer, Mathew. (2005, forthcoming) “Moral and Legal Obligations” in Martin Golding and William Edmundson (eds.) *The Blackwell Guide to the Philosophy of Law and Legal Theory*, Blackwell, Oxford

Kutz, Christopher. (2002) “The Collective Work of Citizenship” in *Legal Theory* 8: 471-494

Kymlicka, Will. (1990) *Contemporary Political Philosophy, An Introduction*, Oxford University Press, Oxford

Lefkowitz, David. (2004) “Legitimate Political Authority and the Duty of Those Subject to It: A Critique of Edmundson” in *Law and Philosophy* 23: 399- 435

Marmor, Andrei. (2001) *Positive Law and Objective Values*, Oxford University Press, Oxford

Markwick, Philip. (2000) "Law and Content-Independent Reasons," in *Oxford Journal of Legal Studies* 20, no. 4: 579 – 596

Markwick, Philip. (2003) "Independent of Content" *Legal Theory* 9: 43-61

McClennen, Edward. (1990) *Rationality and Dynamic Choice: Foundational Explorations*, Cambridge University Press, Cambridge

McClennen, Edward and Scott J. Shapiro. (1998) "Rule-guided behaviour" in *The New Palgrave Dictionary of Economics and the Law*, P. Newman (ed.) (St Martin's Press)

McNaughton, David and Piers Rawling. (1991) "Agent-Relativity and the Doing-Happening Distinction", in *Philosophical Studies* 63: 167 – 185.

Mian, Emran (2002) "The Curious Case of Exclusionary Reasons" in *Canadian Journal of Law and Jurisprudence*, vol. XV: 99 -124

Moore, Michael. (1989) "Authority, Law, and Razian Reasons" in *Southern Californian Law Review*, v 62,

Nagel, Thomas. (1970) *The Possibility of Altruism*, Oxford University Press, Oxford

Nagel, Thomas. (1980) "The Limits of Objectivity" in *The Tanner Lectures on Human Values*, University of Utah Press, Salt Lake City

Nagel, Thomas. (1986) *The View from Nowhere*, Oxford University Press, Oxford

Nozick Robert. (1969) "Coercion" in Sydney Morgenbesser, Patrick Suppes, and Marton White (eds.), *Philosophy, Science and Method* , St. Martin's, New York

Offe, Claus. (1999) "How Can We Trust Our Fellow Citizens," in *Democracy and Trust*, ed. by Mark E. Warren, Cambridge University Press, Cambridge

Overvold, Mark. (1980) "Self-Interest and the Concept of Self-Sacrifice" in *Canadian Journal of Philosophy* 10: 105-118

Parfit, Derek. (1984) *Reasons and Persons*, Oxford University Press, Oxford

Parfit, Derek. (1997) "Reasons and Motivation" in *Proceedings of the Aristotelian Society*, supplementary volume 71: 99-130

Pettit, Philip. (1997) *Three Methods of Ethics: A Debate*, by Marcia Baron, Phillip Pettit and Michael Slote, Blackwell, Oxford

Pink, Thomas. (1996) *The Psychology of Freedom* Cambridge University Press, Cambridge

Raz, Joseph. (1977) "Promises and Obligations" in P.M.S. Hacker and J. Raz (eds) *Law, Morality and Society. Essays in Honour of H.L.A. Hart* Oxford University Press, Oxford

Raz, Joseph. (1978a) "Reasons for Action, Decisions and Norms" in Raz (ed.) *Practical Reasoning* (Oxford, Oxford University Press)

Raz, Joseph. (1978b) (ed.) *Practical Reasoning* (Oxford, Oxford University Press)

Raz, Joseph. (1979) *The Authority of Law*, Clarendon Press, Oxford

Raz, Joseph. (1986) *The Morality of Freedom*, Clarendon Press, Oxford

Raz, Joseph. (1989) "Facing up: A Reply" in *Southern California Law Revie*w vol. 62: 1153 - 1235

Raz, Joseph. (1990a) "Introduction" in "Authority, " Raz, Joseph (ed.), Oxford University Press, Oxford

Raz, Joseph. (1990b) *Practical Reason and Norms*, 2 ed., Princeton University Press, Princeton, New Jersey

Raz, Joseph. (1994a) *Ethics in the Public Domain*, Clarendon Press, Oxford

Raz, Joseph. (1994b) "Moral Change and Social Relativism" in *Social Philosophy and Policy* 11: 139 – 158

Raz, Joseph. (1996) "Liberty and Trust" in Robert George (ed.) *Natural Law, Liberalism and Morality* Oxford University Press, Oxford

Raz, Joseph. (1998) "Disagreement in Politics" in *American Journal of Jurisprudence* 43: 25 - 52

Raz, Joseph. (1999) *Engaging Reason*, Clarendon Press, Oxford

Raz, Joseph. (2001a) "Reasoning with Rules", in *Current Legal Problems* 54: 1-18

Raz, Joseph. (2001b) *Value, Respect and Attachment*, Oxford University Press, Oxford

Raz, Joseph. (2003) "Comments and Responses" in Meyer, Paulson and Pogge (eds.) *Rights, Culture and the Law: Themes from the Legal and Political Philosophy of Joseph Raz*, Oxford University Press, Oxford

Raz, Joseph. (2004) "Incorporation by Law" (2004) in *Legal Theory* 10: 1-17

Regan, Donald. (1989) "Authority and Value. Reflection on Raz's *Morality of Freedom*" in *Southern California Law Review* vol. 62: 995 - 1095

Regan, Donald. (2004) "Why am I my Brother's Keeper?" in *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, edited by Jay Wallace, Michael Smith, Samuel Scheffler and Philip Pettit, Oxford University Press, Oxford

Scanlon, Thomas M. (1999) *What We Owe to Each Other*, Harvard University Press, Cambridge, Mass.

Scanlon, Thomas M. (2004) "Reasons: A Puzzling Duality?" in in *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, edited by Jay Wallace, Michael Smith, Samuel Scheffler and Philip Pettit, Oxford University Press, Oxford

Schauer, Frederick. (1993) *Playing by the Rules. A Philosophical Examination of Rule-Based Decision-Making in Law and in Life*, Clarendon Press, Oxford

Scheffler, Samuel. (1982) *The Rejection of Consequentialism: A Philosophical Investigation of the Considerations Underlying Rival Moral Conceptions*, Clarendon Press, Oxford

Scheffler, Samuel. (ed.) (1988) *Consequentialism and Its Critics* (*Oxford Readings in Philosophy* series), Oxford University Press, Oxford

Searle, John R. (1978): "Prima Facie Obligations", in Joseph Raz (ed.) *Practical Reasoning*, Oxford University Press, Oxford

Shapiro, Ian. (1999) *Democratic Justice*, Yale University Press, New Heaven and London

Shapiro, Scott J. (2000) "The Bad Man and the Internal Point of View" in *The Path of the Law and Its Influence: The Legacy of Oliver Wendell Holmes, Jr*. ed. by Steven Burton, Cambridge University Press, Cambridge

Shapiro, Scott J. (2002a) "Authority" in Jules Coleman and Scott J. Shapiro, (eds.) *Oxford Handbook of Jurisprudence and Legal Philosophy*, Clarendon Press, Oxford

Shapiro, Scott J. (2002b) "Law, Plans and Practical Reason" in *Legal Theory* 8: 387-441

Shapiro, Scott J. (2002c) "Ulysses Rebound" in *Economics and Philosophy* 18: 157-182

Shiner, Roger. (1992) *Norm and Nature. The Movements of Legal Thought*, Clarendon Press, Oxford

Simmons, John A. (2001) *Justification and Legitimacy: Essays on Rights and Obligations*, Cambridge University Press, Cambridge

Soper, Philip. (1989) "Legal Theory and the Claim to Authority" in *Philosophy and Public Affairs* 18: 209-237

Sugden, Robert. (1992) "Inductive Reasoning in Repeated Games" in Reinhard Selten (ed.) *Rational Interaction: Essays in Honour of John C. Harsanyi*, Springer-Verlag Berlin, Heidelberg

Sumner, L.W. (1996) *Welfare, Happiness and Ethics*, Oxford University Press, Oxford

Uslaner, Eric. (1999) "Democracy and Social Capital" in Mark E. Warren (ed.) *Democracy and Trust*, Cambridge University Press, Cambridge

Waldron, Jeremy. (1989) "Autonomy and Perfectionism in Raz's *Morality of Freedom*", in *Southern California Law Review* vol. 62: 1097-1052

Waldron, Jeremy. (1991) "Rights and Majorities. Rousseau Revisitied" in J. Pennock and J.J. Chapman (eds.) *Majorities and Minorities, Nomos* XXIII

Waldron, Jeremy. (1998) "Precommitment and Disagreement" in Alexander, Larry (ed.) *Constitutionalism. Philosophical Foundations*, Cambridge University Press, Cambridge

Waldron, Jeremy. (1999) *The Dignity of Legislation*, Cambridge University Press, Cambridge

Waldron, Jeremy. (2001) *Law and Disagreement*, Oxford University Press, Oxford

Waldron, Jeremy. (2003) “Authority for Officials” in Meyer, Paulson and Pogge (eds.) *Rights, Culture and the Law: Themes from the Legal and Political Philosophy of Joseph Raz*, Oxford University Press, Oxford

Wall, Steven. (1998) *Liberalism, Perfectionism and Restraint*, Cambridge University Press, Cambridge

Wall, Steven. (2003) “Review of Dan-Cohen, Meir. *Harmful Thoughts: Essays on Law, Self and Morality*” *Ethics* 113: 164-167

Waluchow, Wil J. (2000) “Authority and the Practical Difference Thesis: A Defense of Inclusive Legal Positivism” in *Legal Theory* 6: 45-81

Wilkinson, Timothy. (1996) “Dworkin on Well-being and Paternalism,” *Oxford Journal of Legal Studies* 16 no 3: 433-444.

Williams, Bernard. (1981) “Internal and External Reason” in *Moral Luck*, Cambridge University Press, Cambridge

Wolff, Robert Paul. (1970) *In Defense of Anarchism*, Harper and Row, New York

Zimmerman, David. (1981) “Coercive Wage Offers” in *Philosophy and Public Affairs* 10: 121-145