Perfecting the Art of the Possible

A Constraint-Based View of Ideal and Non-Ideal Theory

A thesis submitted to the University of Manchester for the degree of Doctor of Philosophy in the Faculty of Humanities

2015

Brian Carey

School of Social Sciences
# Contents

(i) Abstract 4
(ii) Declaration 5
(iii) Copyright Statement 6
(iv) Acknowledgements 7

**Introduction** 8

1. Ideal and Non-Ideal Theories 12
   1.1 Introduction 12
   1.2 Critical Conceptions 13
   1.3 Alternative Conceptions: The Case for the Defence 30
   1.4 Conclusion 35

2. Theorising (Almost) Without Constraints 36
   2.1 Introduction 36
   2.2 Are A-Type Approaches Coherent? 38
   2.3 Are A-Type Approaches Valuable? 53
   2.4 Why Cohen is an A-Type Theorist 58
   2.5 Conclusion 63

3. The Constraint-Based View 65
   3.1 Introduction 65
   3.2 Describing the Constraint-Based View 66
   3.3 Comparing the Constraint-Based View to the Alternatives 77
   3.4 Conclusion 90

4. The Constraint-Based View and Feasibility 92
   4.1 Introduction 92
   4.2 Feasibility and the Constraint-Based View 95
   4.3 Trade-offs and Fallibility 98
4.4 A ‘Narrowing-Down’ Approach  106
4.5 Conclusion  108

5. Transitional Theory  110
  5.1 Introduction  110
  5.2 Destination Selection  111
  5.3 Route Selection  128
  5.4 Tensions and Trade-Offs  132
  5.5 Conclusion  142

6. The Constraint-Based View in Action: Justice for Future Generations  145
  6.1 Introduction  145
  6.2 Describing the Debate  146
  6.3 Re-describing the Debate  156
  6.4 The advantages of the Constraint-Based View  163
  6.5 A Non-Ideal Approach to the Non-Identity Problem  164
  6.6 Conclusion  172

7. Conclusion  173

8. Bibliography  175

Word Count: 73,934
Abstract

The aim of this thesis is to describe and defend ‘The Constraint-Based View’, which is a particular conception of the nature of, and relationship between, ideal and non-ideal theories of justice. Traditionally, ideal theory is characterised by the assumption of ‘full compliance’, while non-ideal theory is characterised by the assumption of ‘partial compliance’. In other words, ideal theories assume that those for whom the theory is meant to apply will be entirely willing and able to comply with the theory’s requirements, while non-ideal theories do not.

In Chapter 1 of this thesis, I describe and assess this original conception as well as several alternative accounts of ideal and non-ideal theory, in order to offer a broad survey of the existing literature, and to identify the various ways that these conceptions fail to capture fully the relationship between ideal and non-ideal theory.

In Chapter 2, I draw a distinction between two different approaches to theorising about justice. ‘Axiological’ or ‘A-Type’ approaches are characterised by the fact that they include almost no assumptions as inputs to the theory, and that they are not intended to provide action-guiding recommendations as part of the outputs of the theory. In contrast, ‘Practical’ or ‘P-Type’ approaches include additional assumptions as part of their inputs and are intended to form part of a process that ultimately produces action-guiding recommendations.

In Chapter 3, I describe and defend my preferred conception of the relationship between ideal and non-ideal theory – the ‘Constraint-Based View’ (CBV). According to the CBV, there is a spectrum of theoretical approaches that can be more or less ideal, depending on the extent to which they include ‘soft constraints’ as part of the inputs to the theory. Soft constraints are facts about the world that can be changed, in contrast to hard constraints, which are facts about the world that cannot. I argue that this way of thinking about the relationship between ideal and non-ideal theory is more useful than the alternative conceptions considered in Chapter 1.

In Chapter 4, I examine the roles that feasibility considerations should play in ideal and non-ideal theory, from the perspective of the CBV and in Chapter 5, I explain the implications of adopting the CBV for what I call ‘transitional theory’, which is concerned with the transition from the status quo towards a more ideal state of affairs.

In Chapter 6, I offer an example of the CBV in action, by considering its implications for the debate over duties of justice towards future generations.

Brian Carey

‘Perfecting the Art of the Possible: A Constraint-Based View of Ideal and Non-Ideal Theory’

Submitted for the degree of Doctor of Philosophy in the Faculty of Humanities at the University of Manchester

24/03/2015
Declaration

I declare that no portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.
Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the “Copyright”) and he has given the University of Manchester certain rights to use such Copyright, including for administrative purposes.

ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.

iii. The ownership of certain Copyright, patents, designs, trademarks and other intellectual property (the “Intellectual Property”) and any reproductions of copyright works in the thesis, for example graphs and tables (“Reproductions”), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.

iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library’s regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University’s policy on Presentation of Theses.
I expect most people reading this section will have written theses of their own, and so it goes without saying that what appears on the page is never just a product of sitting and reading and writing in isolation, but is shaped in all sorts of ways by one’s supervisors, students, colleagues, family and friends. It’s also true that the work that goes into a thesis doesn’t really begin when one officially enrolls on a PhD course, but much earlier than that – before I arrived at Manchester, I was fortunate to have many lecturers who challenged and encouraged me to keep studying philosophy. Of these, I owe particular thanks to Cara Nine and Adina Preda, my Master’s supervisors at University College Cork and University College Dublin respectively. At Manchester, I received invaluable feedback and encouragement from my supervisors Jonathan Quong, Alan Hamlin, and Miriam Ronzoni, each of whom possesses the impressive and annoying ability to articulate my views better than I can myself.

I am also heavily indebted to the wider academic community at Manchester, in particular to my MANCEPT colleagues for countless hours of stimulating discussion of the topics of this thesis and countless hours of stimulating discussion of anything but this thesis. I owe special thanks in this regard to John Derbyshire, Jack Holme, Stephen Hood, Joe Horton, Sara Kallock, Wen-Chin Lung, Chris Mills, Nicola Mulkeen, Rob Munro, Dean Redfearn, Stephanie Rinaldi, Liam Shields, and Garvan Walshe.

Finally and most importantly, I would not have been able to begin this thesis, let alone finish it, were it not for the support of my parents, Seán and Teresa, and my partner Jonathan. I dedicate this thesis to them, which is the least they deserve for putting up with me.
Introduction

“But what has any of this got to do with the real world?”

- That one student in every introduction to political philosophy class.

Political philosophy makes some people suspicious. It makes them suspicious because they assume that doing political philosophy is supposed to help us solve problems that confront us in ‘the real world’. Despite this ostensible aim, it can often seem as though political philosophers try to deal with the problems of the world by retreating from it.

The worlds that many philosophers seem to be concerned with often bear little resemblance to the one in which we find ourselves. They may seem utopian, abstract, utterly infeasible and ultimately of little apparent relevance to our world and our problems. This can be confusing and especially frustrating for those who become interested in the idea of political philosophy because they want to know what we ought to do here and now, in this world, to deal with what John Rawls himself identified as the “pressing and urgent matters…the things that we are faced with in everyday life” (Rawls, 1999a, p.8).

Some of these worries are relatively easy to deal with, of course – we can explain in a relatively straightforward way why unusual and abstract thought experiments can be useful devices to help us become clearer about our thinking about complex normative issues, for example, or why it is helpful to compartmentalise various normative issues, or work through them in different stages of analysis. But there are deeper worries that persist among some philosophers that manifest themselves as a general scepticism about the usefulness of translating a theory of justice for an idealised world (‘ideal theory’) to a theory of justice for the world we live in (‘non-ideal theory’).

The central aim of this thesis is to address this sort of scepticism, by arguing for a particular conception of the relationship between ideal and non-ideal theory that I call ‘the Constraint-Based View’.
I begin in Chapter 1 by surveying several different ways in which the distinction between ideal and non-ideal theory can be understood. As these views are identified and described, I will argue that most of these views are flawed, in one way or another. The point of this chapter is to serve both as a literature review, and (by identifying the flaws with these views) to make space for my own preferred conception of the relationship between ideal and non-ideal theory.

In Chapter 2, I offer a distinction between two approaches to theorising about justice. I argue that we can conceive of two distinct types of project. Firstly, there are ‘axiological’ or ‘A-Type’ approaches to justice, where the aim is to produce recommendations about how we ought to think about justice as a concept, and the relationship between justice and other values. Secondly, there are ‘practical’ or ‘P-Type’ approaches to justice, where the aim is to produce recommendations about what we ought to do to make the world we live in a more just place. This distinction is important for two reasons. Firstly, it is important because it allows us to see that criticism that is appropriate for a P-Type project will not necessarily be appropriate for an A-Type project. For example, in this chapter I argue that much of the criticism levelled at the work of G.A. Cohen by Colin Farrelly is misguided, because Cohen is engaging in an A-Type approach rather than a P-Type approach. Secondly, this distinction allows me to locate my own project firmly within the context of P-Type approaches to justice.

In Chapter 3, I describe and defend the Constraint-Based View (CBV) of ideal and non-ideal theory. A constraint upon a theory is any fact about the world (or about morality) that a theory takes for granted and which might stand in the way of making the world a more just place. Constraints come in ‘hard’ and ‘soft’ varieties – hard constraints are facts about the world that are assumed to be unalterable. Soft constraints are facts about the world that are assumed to be alterable. Soft constraints come in degrees of ‘softness’ depending on how easy or hard it would be to change the fact in question. According to the CBV, a theory is more ideal to the extent that it does not take relevant soft constraints into account, and more non-ideal to the extent that it does. Having described the CBV in detail, I compare it to the alternative views surveyed in Chapter 1, in order to argue that the CBV presents the most plausible way of thinking about the relationship between ideal and non-ideal theory.

In Chapter 4, I examine the role that feasibility considerations ought to play in developing action-guiding recommendations that aim to make the world a more just place. I begin by
endorsing a view which has recently been developed by Pablo Gilabert and Holly Lawford Smith (2012) which suggests that there are two kinds of feasibility – feasibility in a binary sense, which distinguishes between what is possible and impossible, and feasibility in a scalar sense, which concerns how likely it is that a possible political proposals can be implemented. In this chapter, I am particularly concerned with epistemic constraints with regard to the reliability of our judgments about a proposal’s feasibility. I identify three conditions that determine the reliability of feasibility assessments: the ‘completeness condition’ (which tracks the extent which we are able to identify all relevant constraints), the ‘accuracy condition’ (the extent to which our assessments are themselves accurate with regard to the strength of the constraints in question), and the ‘sufficiency condition’ (the extent to which a political proposal manages to address a sufficient number of relevant constraints in order to achieve the desired state of affairs). I conclude that the capacity to make accurate feasibility assessments is vital for any plausible P-Type approach to justice and that we ought to take steps to improve our capacities to make these kinds of assessments via what I call a ‘narrowing-down approach’.

In Chapter 5, my aim is to show that one of the reasons we have to adopt the CBV is that it can provide us with a relatively clear and straightforward picture of what ‘transitional theory’ looks like, i.e. theory which is designed to help us transition from the status quo to a more ideal state of affairs. The majority of this chapter is dedicated to sketching a series of ‘transitional principles’ which would govern our judgments about the comparative desirability of possible states of affairs, which I call ‘destination comparisons’, and political proposals aimed at achieving a particular state of affairs, which I call ‘route comparisons’. These principles attempt to capture the circumstances under which we are permitted to deviate from the pursuit of an ideal in favour of a next-best alternative. Having described and defended five such principles, I argue that there will be some ‘hard’ cases where we will be unable to make a clear decision with regard to which routes and destinations we ought to select. I conclude that we ought to try to minimize the number of hard cases we are likely to encounter, by trying to maximize our capacities to apply transitional principles in a way that offers clear action-guiding recommendations.

In Chapter 6, which is the final chapter of this thesis, I offer an example of the Constraint-Based View in action, by showing what it might look like if applied to a particular debate in contemporary political philosophy. The example I have chosen concerns our moral obligations to future generations, though the CBV could presumably be applied to a great
many contemporary debates. I have chosen this example in particular because it involves a wide variety of relevant constraints, and it allows me to describe a distinctively non-ideal ‘solution’ to one of the more important problems in the current debate over intergenerational justice. If my account is convincing, this chapter should serve to demonstrate that the CBV can be a useful way to develop action-guiding recommendations to deal directly with some of the most pressing and urgent matters we are faced with in everyday life.
1. Ideal and Non-Ideal Theories

1.1 Introduction

The central aim of this thesis is to describe and defend a particular conception of the relationship between ‘ideal’ and ‘non-ideal’ approaches to questions of justice in political philosophy, which I call the ‘Constraint-Based View’. The central aim of this chapter is to explain why the Constraint-Based View is necessary, by identifying and describing the major views that have emerged from the debate about ideal and non-ideal theory, and by highlighting the ways in which these alternatives offer inadequate accounts of the relationship between ideal and non-ideal approaches to theorising about justice. Having assessed the alternatives, it will remain for me to offer the positive case for the Constraint-Based View in the remainder of this thesis.

While this chapter is intended, in part, to illustrate the many different ways in which ideal and non-ideal approaches can be defined, it will be helpful to begin with some initial, provisional, definitions which will hopefully be sufficiently broad and neutral to serve as a basic starting point, without begging the question against any of the theorists whose definitions will be surveyed below.

Let us say then that ‘ideal theory’ constitutes thinking about what justice would look like in a world that is dissimilar to our own in the sense that there are fewer obstacles in an ideal world that would make justice more difficult to achieve. We can define ‘non-ideal’ theory then, as thinking about what justice would look like in a world that is similar to our own, given all of the various relevant obstacles that might make justice difficult to achieve. Note that these provisional definitions are intentionally neutral with regard to two very important questions: 1) What is the point of an ideal, or a non-ideal approach and 2) What counts as an ‘obstacle’ to the implementation of justice? Once we start to tackle these questions, the definitions of ‘ideal’ and ‘non-ideal’ approaches become more precise than the initial definition above, and we can start to see the ways in which the conceptions to be surveyed below, differ from one another in important respects.

The remainder of this chapter is structured as follows. In Section 1.2, I outline some of the key disagreements between defenders and critics of ideal theory (understood in terms of the
initial definition). In this section I will show that most of these disagreements come down to disputes about the two important questions mentioned above, i.e. what should be the point of ideal and non-ideal approaches to theorizing about justice, and what sorts of things should an ideal or a non-ideal approach take into consideration, when trying to produce conclusions about what we should think about justice, or about what we ought to do to make the world a more just place.

In Section 1.3, I explore the ways in which different answers to these questions have produced different conceptions about how exactly we ought to understand ideal theory, non-ideal theory, and the relationship between the two. My aim here will be to show that (with one important exception), each conception offers an inadequate account of ideal and non-ideal theory, usually by focusing on only one dimension of what turns out to be a multidimensional continuum of theories which can be ideal in some respects, but non-ideal in others. The major exception to this tendency is the view offered by Alan Hamlin and Zofia Stemplowska, who have articulated this “multidimensional continuum” view as an alternative to the other views surveyed in Section 1.2.¹

### 1.2 Critical Conceptions

Perhaps the most powerful intuition which motivates critics of ideal theory is the idea that a properly-constructed theory of justice should (eventually) provide us with action-guiding recommendations which can be applied “here and now” in the non-ideal world. From this perspective, an approach that ultimately fails to provide these recommendations is assumed to be at best irrelevant and at worst inimical to the pursuit of justice. One of ideal theory’s harshest critics expresses this intuition as follows:

> A theory of social justice, and the principles of justice it endorses, must function as an adequate guide for our collective action. A theory of social justice that yields impotent or misguided practical prescriptions is a deficient theory of justice. If the collective aspiration to implement the conclusions of a theory would not result in any noticeable increase in the justness of one’s society, then it fails as a normative theory. (Farrelly, 2007 p.845)

¹ See Hamlin & Stemplowska (2012).
Michael Phillips expresses a similar worry when he claims that the Rawlsian account of ideal and non-ideal theory:

[leaves us] without a method for bridging the gap between a set of political principles and morality appropriate to ideal conditions and a set of political principles and a morality appropriate to the imperfect circumstances within which we live out our lives. (Phillips, 1985 p.551)

Similarly, Amartya Sen, in a paper titled “What do we want from a theory of justice?” criticizes what he calls the “transcendental” approach to justice:

[T]he answers that a transcendental approach to justice gives—or can give—are quite distinct and distant from the type of concerns that engage people in discussions on justice and injustice in the world, for example, iniquities of hunger, illiteracy, torture, arbitrary incarceration, or medical exclusion as particular social features that need remedying. The focus of these engagements tends to be on the ways and means of advancing justice—or reducing injustice—in the world by remedying these inequities, rather than on looking only for the simultaneous fulfillment of the entire cluster of perfectly just societal arrangements demanded by a particular transcendental theory. (Sen, 2006 p.218)

Each quotation above reflects what may seem like a fairly commonsensical assumption: we theorize about normative questions because we want to figure out precisely how we ought to behave— not merely in some abstract, hypothetical thought experiment but in the actual world in which we live. Abstract theorizing may well play a vital role in coming to these conclusions, but to produce a theory which either provides no recommendations for the real world at all, or provides recommendations which turn out to be impotent or misguided, is surely to have made a fatal mistake somewhere along the line.

It is worth noting here that these concerns are shared by John Rawls (the originator of the ideal/non-ideal distinction) as well as his critics, (even if Rawls would not go so far as to claim that guidance for the real world is the only reason to theorize about justice). Having described the problems posed by non-ideal theory as “pressing and urgent matters”, he states quite explicitly in *A Theory of Justice* that “[t]he reason for beginning with ideal theory is that it provides, I believe, the only basis for the systematic grasp of these more pressing problems” (Rawls, 1999a p.8). Rawls clearly agrees to an extent the spirit of the intuition expressed by Farrelly, Phillips and Sen; that an important function of theorizing about justice
is to derive recommendations that can eventually be applied in some form or another in the non-ideal world.²

Where Rawls and these critics diverge, obviously, is that Rawls apparently believes that his ideal/non-ideal approach is capable of delivering the sorts of recommendations required by a healthy theory of justice.³ This marks one of two broad positions which may be taken in opposition to critics of ideal theory- one could begin by accepting the claim that a theory of justice needs to be grounded or connected in some important, recommendation-producing sense, to the real world. From here, it is “simply” a matter of demonstrating that the Rawlsian model (or one’s preferred variant) is up to the task.

An alternative response can be seen in the treatment of ideal theory by theorists such as G.A. Cohen, Andrew Mason, Adam Swift and Zofia Stemplowska, that is, to reject the idea that a normative theory of justice that provides no obvious action-guiding principles for the real world must be fundamentally defective. In a sense, some of these theorists (notably Cohen) attempt to “out-Rawls” Rawls, suggesting that if there is a problem with Rawlsian ideal theory it is not that it operates in a realm that is too remote from considerations about the real world. On the contrary, the problem may be that Rawlsian ideal theory does not go far enough. On this account, there is more to a theory of justice than merely “connecting the dots” between the ideal and the actual in order to generate action-guiding principles. This sort of response attempts to create some space for abstract, conceptual work that need not necessarily be tied to any particular problem that confronts us in the real world. In Chapter 2 of this thesis, I examine these kinds of views in much more detail.

We can consider each of these positions as possible replies to the argument made by Farrelly et al, which can be expressed in a general form as follows:

**P1** – We should reject any method of theorizing about justice that fails to, or is not concerned with, providing us with action-guiding principles in the real world.

**P2** - The Rawlsian method (or any variant that relies heavily upon ideal theory) fails to provide us with action-guiding principles in the real world.

---

² According to Rawls, “If ideal theory is worthy of study, it must be because, as I have conjectured, it is the fundamental part of the theory of justice and essential for the non-ideal part as well.” (Rawls 1999a p. 343)

³ Rawls clearly believes that non-ideal principles are derivable from ideal ones, which is why ideal theory must precede non-ideal: “The ideal conception shows in this case anyway how the non-ideal scheme is to be set up; and this confirms the conjecture that it is ideal theory which is fundamental.” (Rawls 1999a p. 212)
C- We should reject the Rawlsian method (or any similar variant that relies heavily upon ideal theory).

If we consider critics like Farrelly, Phillips, and Sen to be making the above sort of argument, then we can divide the possible defenses of ideal theory into two camps. On the one hand are those (such as Rawls himself) who would reject the second premise but not the first. On the other are theorists such as Cohen, Mason, Swift and Stemplowska who would reject the first (and may or may not reject the second as well). Defeating either premise is, of course, sufficient to defeat the argument, though there will likely be different implications depending on whether one or both premises are held to be at fault.

In order to investigate the premises above, we need to begin by asking questions that are primarily about the nature and the scope of ideal theorizing. We must start by asking: what ideal theories do, and, following whatever conclusions we reach, we must then attempt to determine what it is that we want ideal theories to do. This task carries obvious descriptive and normative components. In descriptive terms, we need to be able to give an account of the various purposes (assuming there is more than one) to which ideal theory can be applied. If one of these purposes is aimed at generating action-guiding principles, for example, (as Rawls and others seem to think) then we need an account of the relationship between the ideal and the non-ideal to demonstrate whether and to what extent the ideal is necessary or useful for generating action-guiding principles which can be applied in non-ideal contexts.

The normative elements of this investigation appear to be twofold: first, if it turns out that ideal theory can be used for multiple purposes, we must ask whether some are to be regarded as being of more fundamental importance than others. This is essentially the main intuition driving many of the critics of ideal theory- even if there are other purposes which ideal theorizing might serve, there is a fundamental value to theory which is focused upon generating action-guiding principles for the real world.⁴

The second normative question that will likely arise concerns how we ought to respond to our preferred descriptive account of the relationship between ideal and non-ideal theory. There would seem to be roughly three broad kinds of responses available, having investigated the relationship between the two and come to a conclusion as to whether ideal theory is

---

⁴ So, Swift says “Those wanting [a theory] to guide political action here and now are making a judgment about the importance of that particular purpose, not a claim about the fundamental character of the discipline” (Swift 2008 p.364)
capable of doing what we want it to do. The first sort of response would hold that the original Rawlsian account is essentially correct, that the critics are somehow in error and that we should proceed accordingly in our theorizing. At the other extreme, if we conclude that Farrelly’s criticism (or something along those lines) is essentially correct, then we might wish to abandon ideal theory altogether if it turns out that it is not only unhelpful but possibly even damaging to the sorts of ends we want to pursue. Between these two extremes, however, lie a range of positions which might consist of fleshed out or modified accounts which include ideal theory but which have been adapted to resist the legitimate criticisms (whatever those might be) of the original Rawlsian account.

Following all of the above, we can summarize the essential questions as follows, each of which corresponds to the first two premises of the critics’ argument, outlined above:

1. What is the point of Ideal Theory?
2. If the point of Ideal Theory is (ultimately) to help us figure out what we should do here and now, is it capable of succeeding in this aim?

For the remainder of this chapter, I will be concerned with presenting the various ways in which theorists have attempted to answer these questions. Having offered a clear and comprehensive map of the theoretical landscape in this chapter, I will offer my own answers to these questions in the chapters that follow in the light of the various positions and arguments that shall be outlined below.

**Conceptual Cartography**

Unfortunately, one of the first unavoidable difficulties one is confronted with in attempting to tackle the debate over ideal and non-ideal theory is the fact that there is no generally agreed-upon interpretation as to what precisely constitutes the distinction between the two. Indeed, the fact that there is what we might think of as a deficit of conceptual clarity in this area may well be one of the few points upon which most theorists agree. With this problem in mind, Hamlin & Stemplowska identify four different approaches with regard to specifying the distinction between ideal and non-ideal theory (Hamlin & Stemplowska, 2012 pp.2-3):

1. Full compliance and non-full compliance.
2. Idealization and abstraction.

3. Fact-sensitivity and fact-insensitivity.

4. Perfect justice (or another value) and local improvement in justice (or another value).

It should be noted that these conceptions, with the exception of the first, are ones that have been pressed in their original form by critics of ideal theory. In response, some defenders of ideal theory have suggested their own alternatives. So, for example, Zofia Stemplowska (2008) criticizes the “fact-sensitivity” account in particular and argues instead that ideal theory should be thought of as theory that fails to issue action-guiding recommendations. In order to get a complete picture of the various ways in which the distinction has been drawn between ideal and non-ideal theory, therefore, it is necessary to examine the conceptions proposed by ideal theory’s defenders as well as those advocated by its critics. In the sections that follow, I shall begin by examining each of the four conceptions listed above, which should offer a good overview of the critics’ conceptions. Having covered these, I shall then review the conceptions that have been put forward by the defenders of ideal theory. For each heading, I associate the particular conception with its most prominent proponent.

1. Full Compliance and non-full/partial compliance [John Rawls]

Traditionally, it is common to conceive of the distinction between ideal and non-ideal theory as being a distinction between theorizing about justice under conditions of ‘full compliance’ versus theorizing about justice under conditions of ‘partial compliance’. Broadly speaking, we can say that the condition of full compliance is satisfied provided that everyone in society understands what justice requires of them, and is fully-motivated to comply with those demands. Partial compliance applies when these conditions are not met. It is important to note, however, that Hamlin and Stemplowska suggest that we should think of the distinction as being between ‘full’ and ‘non-full’ rather than ‘partial’ compliance. This change in terminology allows us to account for the different ways in which the condition of full compliance may fail to be satisfied. For example, Hamlin and Stemplowska point out that

---

5 This characterisation is part of Rawls’ original distinction between ideal and non-ideal theory in *A Theory of Justice* (Rawls 1999a, p.8). (Rawls’ refers to ‘strict compliance’ rather than ‘full compliance’ but the two terms are equivalent).
a society may be partially compliant because only some of its members are fully compliant, or because those who do comply, only comply with some of the demands of justice, or (as seems likely in most real-world cases) a combination of both. Identifying these complexities is important for Hamlin and Stemplowska’s larger project of presenting a spectrum of more ideal and less ideal theories, as opposed to a division between full compliance and partial compliance.

Nevertheless, it is useful at this point to be on the lookout for a pattern which runs through much of the criticism of ideal theory (and which is contrasted by the Hamlin & Stemplowska view), whereby an effort is made to establish two distinct “spheres” of theory- one at the ideal level and one at the non-ideal level.

In offering an overview of the various approaches to the ideal/non-ideal debate, it is helpful to start with the original conception of the distinction between the two. This is to be found in Rawls’ *A Theory of Justice*, where the difference between ideal theory and non-ideal theory is presented in terms of a distinction between full (or ‘strict’) and partial compliance (Rawls, 1999a p.8). A theory is said to assume full compliance if it ascribes certain capacities and motivations to (almost) everyone to whom the theory applies. Essentially, people must know what the theory requires of them, and must be motivated to achieve it. In the context of Rawls’ theory of justice, these are the properties that characterize the people for whom the principles or justice are supposed to apply.⁶

It is not difficult to see why the above might raise the suspicions of critics of ideal theory, especially given the fact that Rawls sets up the ideal/non-ideal approach as a two-stage process; we begin with ideal theory, and then move to non-ideal theory.⁷ A critic might well object that it is more than a little strange to suppose that we will end up with good, action-guiding principles via a method which begins by assuming that most people will understand what justice requires and will be motivated to act accordingly. Rousseau’s aim of taking “men as they are and laws as they might be” (Rousseau, 1782, Book I) seems incompatible with this approach, beginning as it does with a picture of “men as they certainly are not”. Rawls himself commented on Rousseau’s phrase, defining “men as they are” as referring “to

---

⁶ It is important not to confuse the properties that apply to the representative parties in the Original Position, with the properties of the people that the principles of justice are supposed to apply to. While the properties of each may sometimes be similar in cases where we are asked to imagine unrealistic people, ideal theory is concerned with the subjects of the theory, not the properties of hypothetical representatives in the Original Position.

⁷ “The reason for beginning with ideal theory is that it provides, I believe, the only basis for the systematic grasp of these more pressing problems” (Rawls, 1999a p.8)
persons’ moral and psychological natures and how that nature works within the framework of political and social institutions” (Rawls, 1999b p.7). It seems to be the case (or so the objection might go) that the assumption of full-compliance requires us to do so much tinkering with persons’ moral and psychological natures that whatever conclusions we arrive at could never be of any practical use, since whatever recommendations they would offer would be recommendations for beings who are very different to us. In the real world, we can expect that most people won’t have a complete picture of what justice requires of them, at least not without extensive efforts to educate them. Nor is there any guarantee that, even if almost everyone did know what justice required, they would be able and willing to do what was necessary to achieve a just society.

Setting this objection to one side for the moment, there are good reasons to think that the original Rawlsian\(^8\) model of ideal and non-ideal theory as full and partial compliance is too blunt a conception to properly capture the relationship between ideal and non-ideal theory. As Hamlin and Stemplowska point out, the condition of full compliance may fail to be satisfied in cases where an insufficient number of people are fully compliant or in cases where individuals may be compliant to a degree but not to the extent that they count as *fully* compliant. Acknowledging this distinction is particularly important because the principles that apply to one case of partial compliance might be quite different from another. It is one thing, for example, to ask how a fully compliant minority ought to behave in a world where they are outnumbered by non-compliant individuals, and quite another to ask how a collection of “almost” fully compliant individuals might transition to a point where they achieve full compliance (or to a point where they achieve something closer to full compliance).

Even more complexity is uncovered by further distinctions offered by David Estlund and endorsed by Hamlin and Stemplowska between a ‘hopeful’ and ‘hopeless’ theory and between an ‘aspirational’ and a ‘concessive’ theory (Estlund, 2008 pp.263-270). ‘Hopeless’ theories make demands of individuals (and/or institutions) which we have good reason to believe cannot be met (even though compliance may be technically possible). In contrast, a theory is ‘hopeful’ if there is no good reason to think that its demands will not be complied

---

\(^8\) It may be somewhat unfair to describe this as the “original Rawlsian” idea of ideal and non-ideal theory, in the sense that it is not clear that Rawls himself is unaware of the complexities of the distinction between full and non-full compliance. For example, in *A Theory of Justice* (pp.308-312) Rawls considers different kinds of non-compliance in the context of civil disobedience and duties to comply with unjust laws. However, I do believe it is fair to say that the full/partial compliance distinction, in its simplistic form, has been endorsed by a great many, if not the majority of *Rawlsians*, if not by Rawls himself.
with. The distinction between an ‘aspirational’ theory and a ‘concessive’ one turns on whether or not the theory makes demands that take into account the likelihood of compliance by those to whom it applies. That these distinctions are useful I take to be obvious. It seems to follow that if we want to stick to thinking about ideal and non-ideal theory in terms of full and partial compliance (let us call this the “compliance conception”) that we need to find a way of incorporating these distinctions into our view at the very least.

It is worth noting here another important feature of the Rawlsian account of ideal and non-ideal theory beyond the distinction between full and partial compliance. As mentioned briefly above, Rawls envisions the relationship between the two as a kind of two-stage process, with ideal theory preceding non-ideal in order to yield solutions or recommendations for dealing with the “pressing and urgent matters” of the real world. Leif Wenar describes this process as follows:

Completing ideal theory first, Rawls says, yields a systematic understanding of how to reform our non-ideal world, and fixes a vision…of what is the best that can be hoped for. Once ideal theory is completed for a political sub-domain, non-ideal theory can be set out by reference to the ideal. For instance once we find ideal principles for citizens who can be productive members of society over a complete life, we will be better able to frame non-ideal principles for providing health care to citizens with serious illnesses or disabilities. Similarly, once we understand the ideal principles of international relations we will better see how the international community should act toward failed states, as well as toward aggressive states that menace the peace (Wenar, 2008).

This approach suggests a very particular way of thinking about ideal theory and non-ideal theory such that that there is some sharp conceptual division between the ideal and the non-ideal which can allow a theorist to operate purely in one theoretical sphere (while possibly ignoring or rejecting the usefulness of the other). Farrelly, for instance, seems to think that a theorist can more or less do everything that needs to be done, without venturing into (Rawlsian) ideal theory, while defenders of (Rawlsian) ideal theory seem to think that ideal theory can be ‘completed’ before turning to questions of non-ideal theory.

The opposing view holds that we should think of theories of justice as either falling along a spectrum between two extremes (and therefore including both ideal and non-ideal aspects) or that we should think in terms of a “multi-dimensional continuum” as mentioned above, whereby a theory might be ideal in some respects and non-ideal in others. For example, consider that the condition of full compliance requires both a lack of epistemic constraints and a lack of motivational constraints – we have to both know what justice requires, and be
motivated to meet those requirements. A theory may seek to describe which justice looks like when both of these conditions are satisfied, when only one of these conditions is satisfied, or when neither condition is satisfied, depending on the kinds of insights the particular approach is intended to generate. Whichever conception of ideal and non-ideal theory we prefer, be it the compliance conception or some other account (such as the three alternatives I summarize below), we may choose to adopt a dichotomous view of the division between the two or a non-dichotomous view, depending on whether the conception allows us to neatly divide our approaches into completely ideal approaches and completely non-ideal approaches. As I describe the major alternatives to the compliance view below, I hope it will become clear that the dichotomous view is almost certainly unsustainable on any plausible conception of the ideal/non-ideal distinction.

2. Idealization and Abstraction [Charles W. Mills]

Charles W. Mills (2005) offers a criticism of ideal theory that divides ideal approaches from non-ideal by focusing on the distinction between ‘idealization’ and ‘abstraction’ as introduced by Onora O’Neil (1987). Both idealization and abstraction are processes of simplification, intended to allow the core issues of some particular problem to be assessed by removing certain complexities which are likely to arise in the ‘real world’. The crucial difference between the two processes, however, is that abstraction amounts to a kind of ‘bracketing off’ of certain complexities whereas idealization involves simplification through making assumptions that are actually false (Hamlin & Stemplowska, 2012 p.50). According to Mills, because of its reliance on idealization rather than mere abstraction, “the so-called ideal theory more dominant in mainstream ethics is in crucial respects obfuscatory, and can indeed be thought of as in part ideological, in the pejorative sense of a set of group ideas that reflect, and contribute to perpetuating, illicit group privilege” (Mills, 2005 p.166).

Mills singles out several kinds of (what he claims are) idealized assumptions made in ideal theory including: an idealized social ontology (the assumption that the world is comprised of “the abstract and undifferentiated equal atomic individuals of classical liberalism”), idealized capacities (“The human agents as visualized in the theory will also often have completely unrealistic capacities attributed to them”), and an idealized “cognitive sphere” (“A general social transparency will be presumed, with cognitive obstacles minimized as limited to biases of self-interest or the intrinsic difficulties of understanding the world, and little or no
attention paid to the distinctive role of hegemonic ideologies and group-specific experience in distorting our perceptions and conceptions of the social order.”\(^9\)

Let us pause for a moment to consider again the critics’ general argument against ideal theory I presented at the beginning of this Chapter:

- **P1** - We should reject any method of theorizing about justice that fails to, or is not concerned with, providing us with action-guiding principles in the real world.

- **P2** - The Rawlsian method (or any variant that relies heavily upon ideal theory) fails to provide us with action-guiding principles in the real world.

- **C** - We should reject the Rawlsian method (or any similar variant that relies heavily upon ideal theory).

Note that the Mills’ criticism of ideal theory doesn’t quite fit into the above argument. On Mills’ account, the problem with ideal theory is not that it necessarily fails to give us action-guiding principles. Rather, the problem is that the only action-guiding principles it can lead to must be fundamentally deficient since they say nothing about systematic oppression (which has been idealized away before the recommendations have been developed). In fact, we can press this criticism further (as Mills does) and suggest that not only will this approach lead to recommendations that are insufficient for dealing with injustice in the real world, but that following this approach will in fact make it more difficult to achieve our goals in the long run, since the kinds of oppression-related problems it ignores will be allowed to further fester if left unchecked.

Taking this kind of criticism into account, we ought to speak not merely of the need (or rather the assumed need) of a theory to produce recommendations where are merely possible for us to follow. Instead, if a theory’s value is determined by the recommendations it issues, they must be recommendations which are not merely achievable, but also desirable (Stemplowska calls these “AD-recommendations” (Stemplowska, 2008 p.324)). With this in mind, we can reformulate the original general argument against ideal theory”

- **P1** – We should reject any method of theorizing about justice that fails to, or is not concerned with, providing us with action-guiding principles in the real world *that are both achievable and desirable.*

---

P2 - The Rawlsian method (or any variant that relies heavily upon ideal theory) fails to provide us with action-guiding principles in the real world that are both achievable and desirable.

C- We should reject the Rawlsian method (or any similar variant).

Is Mills’ abstraction/idealization distinction a useful one for distinguishing between ideal and non-ideal theory? The major difficulty with this conception seems to lie in determining what precisely counts as an abstraction in some particular context, and what counts as idealization. Hamlin and Stemplowska are skeptical of the usefulness of the distinction in the ideal/non-ideal debate, suggesting that it is not at all clear when one is dealing with complex issues that require simplification, whether a particular simplification ought to count as an abstraction or as an idealization (Hamlin & Stemplowska, 2012 pp.50-51). It is not clear in any case why exactly we should regard abstractions as not involving false assumptions— they entail a process of simplification, and to theorize under the assumption that the situation is more simple than it really is seems to involve a pretty clear falsehood which is nevertheless (one assumes) usually necessary and useful in arriving at a conclusion.

Regardless of whether the distinction is useful or not, Mills’ criticism establishes an important additional hurdle for the defender of ideal theory to overcome; not only must it be shown that action-guiding recommendations can emerge from a process that begins with ideal theory, it must be shown that the recommendations which do emerge are not likely to be deficient or even damaging to the pursuit of justice.

3. Fact-sensitivity and fact-insensitivity [Colin Farrelly]

Colin Farrelly sets up the debate between ideal and non-ideal approaches in terms of “a disagreement over how fact-sensitive a theory of distributive justice ought to be” (Farrelly, 2007 p.844). On the one hand, according to Farrelly, are theorists who “feel inclined to invoke highly abstract hypotheticals when deriving the principles of justice” who stand in opposition to “those political theorists who take seriously real, non-ideal, considerations” (Farrelly, 2007 p.844). While Farrelly speaks about theories functioning “at the level of ideal theory” and thus appears to endorse the dichotomous view of ideal and non-ideal spheres of theorizing mentioned previously, he does acknowledge a distinction between what he calls ‘moderate’ ideal theorists such as Rawls and Ronald Dworkin and the ‘extreme positions’
taken by Mason and Cohen. Farrelly notes that Rawls, even while operating on the ideal level, is concerned with a realistic utopian account. He goes on to concede:

I agree that there are important differences between the extreme and moderate idealized positions. Rawls and Dworkin are not completely blind to non-ideal considerations. In fact, a good deal of their theories is motivated by non-ideal considerations. Rawls, for example, defends a theory of justice that seeks to take seriously what David Hume called the ‘circumstances of justice’. These include the facts of moderate scarcity and limited altruism (Farrelly, 2007 p.846).

Specifically then, Farrelly’s distinction is not merely between fact-sensitive and fact-insensitive theories, but between theories that are sufficiently fact-sensitive and those that are not. Farrelly’s main target are the ‘moderately’ ideal theorists who he argues are themselves leaning too far in the direction of the ideal, but it is not at all clear that we should accept his positioning of Cohen and Mason as being so committed to ‘fact free’ theory that they create an unbridgeable gulf between the ideal and the non-ideal. Adam Swift has accused Farrelly of setting up a false opposition in claiming a contrast between Mason and Cohen who believe (according to Farrelly) that ‘the fundamental principles of justice are logically independent of issues of feasibility and questions about human nature’ and theorists “who believe that normative theorizing must be integrated with an appreciation of the empirical realities of one’s society” (Farrelly, 2007 p.844). Swift responds:

But the two views described here do not “contrast strongly.” One can hold both that “the fundamental principles of justice are logically independent of issues of feasibility and questions about human nature” and (as long as the “must” is read as a “should” rather than as claiming conceptual necessity) that “normative theorizing must be integrated with an appreciation of the empirical realities of one’s society”… But as long as one was clear about what exactly one was doing at each stage of the argument—and one would need to attend carefully to what one was calling a “principle of justice”—there is no reason why one could not both “feel inclined to invoke highly abstract hypotheticals when deriving the principles of justice” and “take seriously real, non-ideal considerations” (Swift, 2008 p.371).

Even if we are willing to accept Farrelly’s account for the sake of argument, there appears to be a crucial piece of the puzzle that is missing; if what we are concerned with is whether a theory is sufficiently fact-sensitive, how are we to judge what counts as “sufficient” in this context and what does not? To be fair to Farrelly, he acknowledges that this is a question that must be addressed, but puts it beyond the scope of his paper. Nevertheless, if this conception is to be useful (call this the ‘fact-sensitivity conception’), it is surely the case that it can only be useful if we can determine a way of distinguishing between theories that are sufficiently
fact-sensitive and those that are not. For each particular theory, there will presumably be a range of facts it must take into account such that if it fails to do so it will provide recommendations that are “impotent or misguided”. Farrelly hints at what sorts of facts might count when he lists some of the constraints which non-ideal theory must “take seriously” (Farrelly 2007 p.847):

- Non-compliance
- Unfavorable historical, social or economic conditions
- Indeterminacy
- Fallibility
- Disagreement
- Human vulnerability
- Human nature
- Problems of institutional design, etc….

It’s worth noting here that there may be a danger for Farrelly that his own approach is ultimately self-defeating, if it turns out that we cannot find a reliable way of distinguishing which facts a theory needs to take into account from those it does not. This is because Farrelly’s dismissal of ideal theory rests upon the core claim that a normative theory that is ‘impotent’ is a theory that is useless. It seems possible that one might, with minimal modification, use a similar line of argument to suggest that Farrelly’s own position ought to but fails to live up to these very standards, since it establishes a distinction in theory which we are unable to make in practice (it should be re-emphasized here that Farrelly does note this further challenge for his position, though he places it beyond the scope of his initial criticism). For example, suppose that for our theory to issue recommendations that are not impotent, it must be the case that we take particular facts about human motivational capacities into account. If there are certain kinds of demands which human beings, owing to some fundamental and permanent feature of human psychology, are unable to meet, then it looks like Farrelly’s view suggests these facts must be taken into account. However, this presupposes a capacity (or a capacity to develop a capacity) to determine which features of human psychology ought to be taken into account, which features are permanent, and which might be modified. A fully-developed version of Farrelly’s preferred approach, therefore, must be able to identify the necessary and sufficient set of facts which are required in order to make a theory sufficiently fact sensitive.
4. The distinction between a theory of perfect justice (or another value) and a theory of local improvement in justice (or another value). [Amartya Sen]

In “What do we want from a theory of justice?” (2006) and The Idea of Justice (2009), Amartya Sen criticizes what he calls the ‘transcendental institutionalist’ approach to justice, as exemplified by Rawls (though Sen claims this approach can also be seen in the work of Hobbes, Locke, Rousseau and Kant), and contrasts this with the idea of making ‘comparative judgments’ between different non-ideal states of affairs. The former approach, Sen claims, is neither necessary nor sufficient to generate action-guiding recommendations for the world we live in. This idea seems to stand in stark contrast to Rawls’ approach, whereby ideal theory is assumed to be a necessary first step in tackling the “pressing problems” of the non-ideal world. While it should be noted that Sen does not explicitly map his transcendental/comparative distinction onto the ideal/non-ideal one, some theorists regard it as uncontroversial to treat the two as being more or less analogous\(^{10}\) with at least one notable exception (see below).

According to Sen, transcendental approaches begin with the question “What is a just society?” (Sen, 2006 p.216) This leads to a focus on “perfectly just societal arrangements”. In contrast, the comparative approach is not concerned with what a perfectly just society would look like, but rather with ranking alternative non-ideal arrangements in terms of one being more just or less just than the other. Again, we can note that Sen’s account is explicitly dichotomous: “The transcendental and comparative approaches are quite distinct, and as will be presently discussed, neither approach, in general, subsumes or entails the other.” The core idea behind Sen’s criticism of the transcendental approach is that knowledge about what a perfectly just society would look like is neither necessary nor sufficient for tackling injustice in the non-ideal world. Laura Valentini identifies three key components of Sen’s criticism, offering the following claims (Valentini, 2011 pp.297-315):

---

\(^{10}\) On this point, see Hamlin & Stemplowska (2012) and Valentini (2007)
(1) Transcendental institutionalism is neither necessary nor sufficient to carry out comparative judgments of justice.

(2) Transcendental institutionalism is inherently parochial.

(3) Transcendental institutionalism is inflexible.

The above objections seem to contain threads that we have already seen in the criticism of Farrelly and Mills. Recall that Farrelly’s fact-sensitivity conception has as its central claim the idea that ideal theory is so detached from important facts about the real world that it is unable to issue action-guiding recommendations and is rendered practically impotent. This seems to align with the claim that transcendental institutionalism is neither necessary nor sufficient to carry out comparative judgments of justice. The second and third objections seem to align with those of Mills—recall that Mills’ major concern with the ‘abstractions’ of ideal theory was not necessarily that they did not lead to action-guiding recommendations but rather that whatever recommendations they might produce would be so deficient as to be positively counterproductive to the pursuit of justice.

Laura Valentini has offered a defense of the ‘Rawlsian Paradigm’ in response to Sen’s criticisms. In response to the first claim, Valentini’s strategy is to deny first of all the idea that the only purpose a theory of justice should serve is to enable us to make comparative judgments. This is an important point that challenges the first premise of the general critics’ argument I outlined at the beginning of this chapter (i.e. that “we should reject any method of theorizing about justice that fails to, or is not concerned with, providing us with action-guiding principles in the real world that are achievable and desirable”). I shall say no more of this objection here, as it is one I will devote substantial attention to in the following chapter. Valentini goes further than this claim, however, and also suggests that, at least in some cases, the transcendental approach is quite capable of helping us to make comparative judgments:

[I]t seems undeniable that principles like the ones Rawls proposes do take us a fairly long way towards the possibility of making some comparative judgments of justice and injustice. For instance, the principles clearly establish (if one accepts them) that violations of basic liberties are more serious than violations of fair equality of opportunity which, in turn, are more serious than violations of the difference principle (Valentini, 2010 p. 10).

Sen’s second major objection concerns the scope of the Rawlsian method which Sen regards as being inappropriately limited especially when we come to think about questions of
global justice. Valentini’s response here is not to deny that there are restrictions upon the Rawlsian model (in its original form, for example, parties in the Original Position are members of the same society) but to point out that these restrictions stem from a moral, rather than a methodological assumption about the extent to which justice applies to people who stand in certain kinds of morally significant relationships to one another. While Valentini’s reply here does not diffuse Sen’s objections as such, it does suggest that Sen is targeting the wrong aspect of Rawls’ theory.

Sen’s third criticism concerns the supposed “inflexibility” of the Rawlsian account, as Valentini explains:

[Rawls’ theory] aims to be complete, to establish what perfect justice requires, while in fact a ‘partial’– albeit abstract and general – ideal of justice is all we need and can plausibly reach. Second, the principles Rawls defends do not seem to be open to revision. They indicate what perfect justice requires, and are thus as demanding as they are inflexible. This makes Rawls’s theory of justice unable to offer a basis for a fruitful public discussion, and reduces both its theoretical and practical appeal (Valentini, 2011 p.312).

Valentini’s reply to this is to point to Rawls’ later work, where she suggests that Rawls acknowledges the above sort of criticism, to an extent, and goes some way towards making his theory less ‘optimistic’ and thus less vulnerable to such an objection. So for example, Valentini notes that in Political Liberalism Rawls “acknowledged that there is a plurality of reasonable liberal conceptions of justice, which may be less distributively egalitarian than his own ‘justice as fairness’” (Valentini, 2010 p.12). Valentini goes on to point to the Rawlsian idea of “reflective equilibrium” as evidence that the Rawlsian model is open to revision. Reflective equilibrium essentially involves a process of investigating (or reflecting upon) one’s deeply held moral convictions and subjecting them to intellectual scrutiny in order to (hopefully) achieve some kind of overall coherence. Crucially, however, while we may orient ourselves via what we feel are our most fundamental intuitions, there is always an element of provisionality at play, such that we ought to be open to revising our considered judgments as part of the deliberative/reflective process.\(^\text{11}\)

\(^\text{11}\) It should be noted that Hamlin & Stemplowska have argued that it is a mistake to map Sen’s transcendental/comparative model onto the ideal/non-ideal distinction. They argue that the transcendental/comparative distinction only seems to align with the ideal/non-ideal distinction if one adopts additional pre-theoretical assumptions about ‘localness’ and ‘realism’. ‘Localness’, for instance, is necessary in order to fix the bases of comparison at the non-ideal level, given that it is perfectly possible to compare different states of affairs at any level of idealness. See Hamlin, A. & Stemplowska, Z., 2012, p.52.
1.3 - Alternative Conceptions: The Case for the Defense

In defending ideal theory from critics such as Farrelly, Mills and Sen, some theorists have set out their own preferred conceptions of the distinction between ideal and non-ideal theory. In this section, I outline several of these conceptions.


In “What’s Ideal about Ideal Theory?” Zofia Stemplowska begins by offering a kind of conceptual anatomy of theories in general. According to Stemplowska, any theory can be thought of as containing a particular ‘internal structure’ consisting of the inputs of a theory, the outputs of a theory, and the rules of derivation which link one to the other (Stemplowska, 2008 p.323). If this is correct, it suggests that it is possible to classify theories in terms of their inputs, outputs, or rules of derivation. So, for example, one way of distinguishing normative theories from descriptive ones would be to look at the inputs of the theories and to note that all and only normative theories contain at least one normative principle as part of their inputs. Stemplowska suggests that, for example, Farrelly’s account (the fact-sensitivity conception) can be explained as a criticism that focuses on particular inputs of a theory and divides those that contain the proper facts as part of their inputs from those that do not. Of course, as Stemplowska notes, Farrelly’s reason for distinguishing between theories based on their inputs rests on the idea that particular inputs will necessary lead to particular outputs (or that a lack of particular inputs will make the theory unable to generate the necessary outputs). Given that Farrelly claims that normative theories that do not issue us with guidance for action in the real world are fundamentally flawed, Stemplowska labels the sorts of outputs demanded by Farrelly as “AD-Recommendations” (Stemplowska, 2008 p. 324), that is, recommendations for action that are achievable and desirable in the non-ideal world. With this distinction established, Stemplowska moves on to defend ideal theory that does not produce AD-Recommendations.

Stemplowska’s account of the internal structure of a theory and her distinction between theories based on whether or not they produce AD-Recommendations is particularly important because it allows us to uncover another distinction between theories, which turns
out to be extremely important in answering one of the two central questions posed above, namely, “what is the point of ideal theory?”

This additional distinction is between theories that work as they are supposed to, and those that do not. Stemplowska’s conception of ideal theory sees it as playing a very particular kind of role:

[Ideal theory] helps us to see our principles and problems more clearly, it ensures that even when we are not motivated to do what can be required of us we are not thereby let off the hook, and it allows us to uncover more of what we value and should therefore pursue. Ideal theory therefore uncovers, clarifies, and safeguards our normative commitments (Stemplowska, 2012 p.339).

Notice that this account does not seem to fit at all with Farrelly’s claims about the point of ideal theory. For Farrelly, normative theories are supposed to produce particular outcomes (in this case, AD-Recommendations), and they fail as normative theories if they fail to produce them. However, Stemplowska’s Ideal Theory is not malfunctioning by not producing AD-Recommendations - the outputs it is intended to generate involve matters of conceptual clarity, not action-guiding recommendations.

We can note here that there are two very different ways a theory might be criticized in general: We could criticize a theory because it does not function as it is supposed to- a theory is intended to generate particular outputs but fails, so the theory must be scrapped or modified in order to produce what is intended. Or we might admit that a theory functions perfectly fine- the inputs, rules of derivation and outputs all work as they are supposed to, but we might criticize the original purpose behind the theory itself.

Farrelly’s central criticism of ideal theory explicitly assumes that normative theorizing about justice has a central purpose (offering guidance in the non-ideal world), and that ideal theory, in failing to do this, is therefore defective. So he seems to be making the first sort of criticism outlined in the preceding paragraph. If we recall how Rawls envisions ideal theory, as step one of an approach leading to the resolution of the “pressing problems” of the non-ideal world, it is not difficult to see why Farrelly takes this approach. Note, however, that Farrelly places theorists like Andrew Mason and G.A. Cohen on a spectrum with the more “moderate” ideal theorists like Rawls and Dworkin. This suggests that Farrelly considers all
of these theorists to be engaged in the same sort of theoretical project, with someone like Cohen offering a theory that is simply worse at achieving the appropriate outputs than that offered by Rawls.

But suppose that Cohen, for example, is not engaged in the same kind of project as Rawls. Suppose that Rawls is aiming, as Farrelly supposes, to eventually derive action-guiding principles whereas Cohen is aimed at achieving conceptual clarity as to the nature of a concept like justice along the lines of ideal theory as described by Stemplowska (2008). While Rawls is still potentially vulnerable to Farrelly’s criticism, Cohen and other ideal theorists can simply sidestep the entire thrust of Farrelly’s attack simply by denying the first premise of the general argument outlined at the beginning of this chapter, i.e. that we should reject any method of theorizing about justice that fails to, or is not concerned with, providing us with action-guiding principles in the real world.

A theorist like Cohen, then, could simply reply to Farrelly that he is engaged in a project that seeks to produce conceptual clarity, rather than action-guiding recommendations. It is surely not a valid criticism of the kind of theory he produces then, to object on the grounds that it does not produce something it is not intended to produce. In order to mount a successful attack on Cohen, Farrelly must dispute the claim that ideal theory (in the sense Stemplowska outlines) is worth doing in the first place. However, this sort of criticism cannot rest on the theory’s failure to produce action-guiding recommendations; it must instead establish that there is no value to producing a normative theory that is not aimed at producing such outputs in the first place.

I say their “brand” of ideal theory, because I am not convinced that it is better to think of any theory that aims to issue AD-Recommendations as non-ideal theory. Rawls does seem to think that the point of his project is ultimately to arrive at a point where AD-Recommendations may be produced. Stemplowska seems to acknowledge this in suggesting that “complex normative theories, such as Rawls’ theory of justice, are likely to contain within themselves both ideal and non-ideal theory” (Stemplowska, 2008 p.339).

In any case, whether we think of theory which aims for conceptual clarity rather than the production of AD-Recommendations as ideal theory, or some particular species of ideal theory, the fact remains that many of the criticisms of ideal theory are likely to miss the mark when it comes to those theorists who are explicitly concerned with outputs which may have
nothing to do with guiding our actions in the non-ideal world. Unless we are sceptical of the existence of such projects (which seems absurd), critics of ideal theory must deploy arguments capable of questioning the worth of these projects, rather than focusing on outputs that were (at least in some cases) never intended to be produced.

6. ‘Good’ Ideal Theory and ‘Bad’ Ideal Theory [Laura Valentini]

Laura Valentini’s solution to what she calls the ‘guidance critique’ (i.e. the claim that ideal theory ought to, but does not, produce good, action-guiding principles) is to make a distinction, not between ideal and non-ideal theory (here she endorses the idea that ideal theory is “any theory constructed under false, that is, idealized, assumptions, which make social reality appear significantly ‘simpler and better’ than it actually is,” (Valentini, 2009 p. 332) but to draw a second distinction between ‘good’ ideal theory and ‘bad’.) According to Valentini, “whether an ideal theory is good or bad depends on whether it entails a false, idealized, account of the subject to which it is meant to apply.” (Valentini, 2009 p. 355) [Original emphasis].

The first point to note is Valentini’s reference to the “subject” of a theory. The suggestion here is that ideal and non-ideal approaches may be applied to a variety of subjects, from Rawls and Dworkin’s original target (what would a ‘well-ordered’ closed society look like?) to subjects like justice between societies (Rawls’ Law of Peoples12) or justice between liberal majorities and minorities (Kymlicka’s Multicultural Citizenship13). Valentini argues that the distinction between good and bad ideal theory turns on whether or not the theory’s idealized assumptions are “temporary” such that they apply as the theory is being constructed but can then be addressed as the theory expands to take non-ideal problems into account, or “permanent” such that they are integral to the very conception of the subject of the theory. So, for example, Valentini claims that Kymlicka’s theory is “bad” ideal theory because Kymlicka resists applying it to the real world, and that Rawls’ international theory is bad ideal theory because Rawls refuses to drop the assumption that states will be well-ordered even when considering non-ideal problems that arise between states.

In contrast, Valentini suggests that while Rawls’ and Dworkin’s original theories about justice make certain omissions that are necessary to achieve conceptual clarity but which do not serve to incapacitate the theory from subsequent application to non-ideal cases. Such theories establish basic foundational premises onto which theorists can build the additional problems posed by the non-ideal world. Valentini takes gender injustice as an example and argues (following Okin\textsuperscript{14}) that Rawls’ theory is sufficiently flexible to accommodate non-ideal problems:

Rawls suggests that if a major cause of women’s social disadvantage turned out to be the lack of material recognition for their contribution to society’s continued existence through raising children and caring for the family, their contribution should be officially recognized, for instance, by promulgating laws entitling them to half of their husbands’ earnings during the marriage.\textsuperscript{59} From society’s viewpoint, by taking care of children, and ensuring that they become fully capable citizens, women’s activity is just as valuable as that of their partners and therefore deserves recognition (Valentini, 2009 p.344).

If we consider Valentini’s approach in response to the general critics’ argument presented at the beginning of this chapter, we can note a contrast with the sort of approach developed by Stemplowska, Swift, and Hamlin. Whereas Stemplowska and Swift, for example, challenge the premise that the only sort of ideal theory worth having is one which produces AD-Recommendations, Valentini appears to concede more to the critics, accepting the notion that ideal theory should issue AD-recommendations, but rejecting the claim that Rawls (and Dworkin) are incapable of meeting this challenge.

It is not clear to what extent Valentini’s response is actually incompatible with Stemplowska \textit{et al}, however. If we adopt the distinction suggested in the previous section between theories which do “what they are supposed to” and those which do not, then it is possible to bracket Cohen-type theorists off to one side, and then to distinguish between theories that aim to produce AD-Recommendations and succeed, and theories that aim to produce AD-Recommendations and fail. With this distinction in mind, whether Valentini’s good/bad distinction is appropriate will depend on whether the theories she labels as “bad” ideal theories are actually intended to (ultimately) be applicable to the non-ideal world.

1.4 Conclusion

In this chapter, my aim has been to offer an overview of the different ways that we can conceptualise ideal and non-ideal theory, and to make space for an account of my preferred conception by describing the limitations of the alternatives. I will return to these limitations in more detail in Chapter 3, by contrasting the approaches described above with my own. For the moment, it should suffice to have shown that there is considerable controversy over what we ought to think is the point of theorising about justice, how we ought to draw the line between ideal and non-ideal theory, or even whether we ought to draw a clear line between the two in the first place. Part of my task in offering what I believe to be the best account of ideal and non-ideal theory, will be to answer these questions in a way that can incorporate the most useful aspects of the above views, while avoiding their shortcomings or expanding upon the insights they contain.

In the next chapter, I begin by investigating the first of these questions – what is the purpose of theorising about justice? My answer will be that there can be more than one purpose, and that ideal and non-ideal theory will be appropriate for some of these purposes, but not others.
2. Theorizing (almost) Without Constraints

2.1 Introduction

A ‘constraint’ upon a theory is an assumption which is included as part of its inputs that serves to limit the kinds of recommendations that the theory can produce. Constraints pick out particular facts to which the theory will be sensitive – these facts can be facts about the rules of logic, facts about the physical world, or they may be normative claims that the theorist wants to take for granted for the purposes of their theory. A constraint is a ‘hard’ constraint if the assumption is that it represents a fixed and unalterable fact about the world. A constraint is a ‘soft’ constraint if the assumption is that it is not.

We can distinguish between different approaches to theorizing according to the kinds of constraints that are included as part of the inputs to the theory. Additionally, when assessing any particular theory, we ought to take into account whether or not the theory functions as intended. This latter distinction is compatible with, but separate from, constraint-based distinctions: an analysis of the constraints a theory uses as part of its inputs will tell us what sort of a theory we are dealing with, and an assessment of whether or not the theory functions as intended will tell us whether or not the theory is in one important sense ‘good’ or ‘bad.’

A theory ‘functions as intended’ if it produces the outputs that it is intended to produce.

In this chapter I will focus on approaches to theorizing that (1) do not include as part of their inputs any sorts of constraints with the exception of basic assumptions about the rules of logic, which I take to be hard constraints, and (2) whose intended function is not to produce action-guiding recommendations, but to offer us a deeper understanding of the nature of, and relationship between, the principles and values we affirm (such as ‘justice’, ‘fairness’, ‘equality’, ‘welfare’ etc.) and the reasons we have for affirming these values in particular.

I will defend the claim that the approach of G.A. Cohen typifies this kind of theorizing, which I shall refer to in this chapter as axiological, or ‘A-Type’ theorizing. Both defining characteristics of A-Type theories contrast with approaches like those of John Rawls and Ronald Dworkin, which I shall call practical or ‘P-Type’ theories. First, P-Type theories will

---

15 It should be noted that a ‘bad’ theory is not necessarily a useless one. It is possible that a theory may not function as intended but still produce useful results or insights, or that a theory may function precisely as intended, but produce results that are trivial or useless.
include (at a minimum) consideration of a broad range of hard constraints (such as hard epistemic and motivational constraints), whereas A-Type theories will not consider any constraints except those of logical possibility. Secondly, whereas A-Type theories are intended to provide a description of the nature of and/or relationship between our fundamental values, P-Type theories are intended to supply theoretical foundations that will ultimately lead to action-guiding recommendations which are achievable and desirable in the real world.

In this chapter I aim to show 1) that this A-Type/P-Type distinction is coherent, 2) that Cohen’s approach is an example of A-Type theorizing and 3) that the arguments made by Colin Farrelly against what he calls ‘ideal theory’ are, as a result, only applicable to P-Type theorists.

I will defend each of these claims below, but even if I am successful, Farrelly still has the option of accepting each claim but denying the value of A-Type theories themselves. In anticipation of this sort of response, I will begin in Section 2.2 by defending the coherency of the A-Type/P-Type distinction before turning to defend the value of A-Type theorizing in general in Section 2.3. The idea here is to shut off any possible escape route for Farrelly before turning to my final task in Section 2.4, which will be to establish that Cohen’s approach does in fact fall into the category of A-Type theorizing. If I succeed at each step, we can conclude that Farrelly’s criticism applies only to a subset of those theories towards which his arguments were originally aimed (i.e. a certain subset of P-Type theories). It will remain for me to defend the value of these sorts of P-Type theories in Chapter 3 and to explain what further relationship, if any, holds between A-Type and P-Type approaches.

For the purposes of this chapter, since my aim is to contrast these two different approaches with Farrelly’s spectrum of ‘ideal theory’, I shall not be concerned here with non-ideal theorizing which goes beyond hard-constraints-only versions of P-Type theorizing to include soft as well as hard constraints. The view I will ultimately argue for will suggest an alternate spectrum-based view, consisting only of P-Type theories of different strengths and comprehensiveness, in contrast to Farrelly’s view, which characterizes A-Type theories as merely being more extreme/more abstract versions of P-Type theories. My view, which I take to be consistent with Hamlin & Stemplowska’s (2012) distinction between ideal theory and
“the theory of ideals” places P-Type approaches on a spectrum, but divides these from A-Type projects which, as I will argue below, belong in their own distinct category.\(^\text{16}\)

### 2.2 Are A-Type Approaches Coherent?

Zofia Stemplowska writes:

> The main purpose of theories that do not aim at issuing any recommendations seems to be simply to gain a clearer understanding of our values and principles. I find it hard to imagine that anyone might object to such theorizing or that there is much mileage in debating whether this type of analysis should merit the name of normative theory (Stemplowska, 2008 p.330).

I take the above to be a characterization of the core purpose of A-Type theorizing. In a later paper with Alan Hamlin, this approach seems to be broadly similar to what they term ‘the theory of ideals’:

> In the theory of ideals the purpose is to identify, elucidate and clarify the nature of an ideal or ideals (we will call this ‘specifying ideals’). More precisely, there are two component elements to the theory of ideals, one devoted to the identification and explication of individual ideals or principles (equality, liberty, etc.), the other devoted to the issues arising from the multiplicity of ideals or principles (issues of commensurability, priority, trade-off etc.) (Hamlin & Stemplowska, 2012 pp.8-9).

It is not disputed that theorizing about justice must involve recommendations about what we should think when we think about what justice is. Rather, the central theoretical dispute\(^\text{17}\) centres upon whether or not it is possible to think about justice in a way that does not require us to say anything about what we must do in response to what we must think. In other words, is there something fundamentally defective in an approach that seeks to describe what justice is, without telling us anything about how we ought to act, here and now?

A theory aimed at providing conceptual clarification\(^\text{18}\) whose only recommendations concern how we ought to think about some concept or the relationship between some

---

\(^{16}\) Hamlin & Stemplowska’s view conceives of the theory of ideals as being distinct from the ideal and non-ideal theories, whereas on my view both A-Type and P-Type theories are conceived as being forms of ideal theory. I use my own terminology rather than borrowing theirs because I want to allow for the possibility that A-Type approaches lie within ideal theory while still being categorically distinct from P-Type approaches.

\(^{17}\) As opposed to the more practical dispute concerning the allocation of intellectual resources, which I shall deal with below.

\(^{18}\) When I refer to ‘conceptual clarification’ in this chapter, I have in mind a broad range of claims and arguments aimed not only at offering a descriptive account of the nature of and relationship between the ideals
concepts is obviously useful, in general. This cannot be what Farrelly is objecting to, given that his objections themselves are intended to provide recommendations about how we ought to think about justice, and the usefulness/coherency of certain types of theorizing. Farrelly quotes Andrew Mason who asks, “What reason do we have for thinking that any adequate analysis of an ideal such as justice must be conducted in the light of an investigation of what is feasible?” (Mason, 2004, p.255) and replies that he believes:

...there is some conceptual incoherence involved in saying ‘This is what justice involves, but there is no way it could be implemented’ (Mason, 2004, p. 255). This incoherence stems from the fact that a theory of social justice, and the principles of justice it endorses, must function as an adequate guide for our collective action. (Farrelly, 2007, p.845)

Let’s call this the ‘Context-Embedded’ conception of justice; i.e. the idea that there is something intrinsic to this thing we call ‘justice’ such that any complete and coherent analysis of the concept must embed that concept within a particular context, specifically, within the context of ‘the real world’. On this view, justice is all about what we should do here, or what they should do over there, etc. It is inherently action-guiding and must contain or entail some prescriptive element.

This view is apparently compatible with the claim that a theory need not produce recommendations about how we ought to act in order for it to be coherent and useful in general. The claim here, however, is that justice is a concept that cannot properly be described without taking, for example, feasibility considerations into account. If we imagine ‘justice’ as a list of instructions telling us what we ought to do, describing what justice is necessarily requires us to investigate what it is possible for us to do in the first place, which requires that our theory be sensitive to the real world in a way that goes beyond merely taking logical possibility into account. The central dispute, of course, is whether it is correct to imagine justice in this sort of way.

Mason’s position, in contrast, holds that we can separate the analysis of the ideal of justice from considerations about the implementation of that ideal (or feasibility considerations). Call this the Context-Free conception of justice. On this view, justice is something we can fully and coherently describe without having to consider whether and to what extent it would be possible to implement the ideal in the real world. On this view, we can make a distinction between two kinds of normative theory: those theories that merely describe normative that we affirm, but also claims which can include arguments for why we ought to affirm certain values or ideals over others.
concepts, and those theories that produce or imply normative conclusions. In the first case, the subject of the theory is a normative concept (such as freedom or equality, for example), whereas in the second case, ‘normative’ refers to a property of the theory itself, rather than its subject. A theory that recommends that we must refrain from causing unnecessary suffering would be an example of this kind of normative theory.

This distinction seems perfectly plausible when made about non-normative topics. A theory of coherency, for example, is clearly different from a coherent theory. It is less clear whether the distinction holds in cases involving normative subjects, however. Certainly, when it comes to the concept of justice, Farrelly’s central claim seems to be that it does not; i.e. that it is simply incoherent to attempt to merely describe what justice is as a normative concept, without being sensitive to facts about the real world and without producing a theory that necessarily contains action-guiding content. In contrast, consider the following passage from Mason:

…a distinction…needs to be drawn between a concrete use of the term 'ought', which is governed by what is possible and which licenses the inference from 'ought' to 'can', and an abstract use of the term, which brackets issues to do with what is possible and which does not license this inference. It is the abstract use of 'ought' that is involved in principles of justice, and indeed in other general moral principles, such as the principle that people should keep their promises. This abstract sense may also be used in particular judgements when these are intended to draw attention to some morally relevant feature of a situation, rather than express an all things considered judgement about what should be done. When this is the case, 'A ought to do X' means something like, 'A has a moral reason to do X'. (Mason, 2004, p. 257)

We can conceive of the central dispute, given the above, as a dispute over whether or not a complete and coherent theory of justice involves a normative component which is purely ‘abstract’ in the sense Mason describes above, or whether (as Farrelly insists) a description of justice must involve an all-things-considered account of what we ought to do. I do not mean to suggest that these two options exhaust all possible conceptions of justice; it may be the case, for example, that both Farrelly and Mason’s positions capture different but complimentary aspects of justice that can be integrated into a single coherent conception.

There may be a similarity between this dispute and the debate over what it means to say that some person “has a (moral) reason” to do something in describing some particular situation involving normative action.¹⁹ Suppose, for example, that my neighbour is having a seizure at this very moment, but that I am unaware of this. If we ask a question like “are there any facts about the world which would provide me with a reason to aid my neighbour?” the

answer is clearly yes. From this perspective (which is essentially a kind of third-person omniscient one), there is a reason for me to aid my neighbour, and I may be said to ‘have’ this reason in this sort of way.

We might think it’s incoherent to say, “I have a reason to aid my neighbour” in this case, if we equate “X has a reason to φ” with the idea that “X could φ and X knows that he or she ought to φ”. It would seem to follow, if we make this equation, that the claim that I have a reason to aid my neighbour is false or incoherent in this case because there is no way for me to offer aid if I don’t know that aid is required (there may be additional complications depending on whether or not I am at fault for my ignorance, but I shall set those questions aside here). For me to be blameworthy, I must have been able to act, I could not act without knowing that my neighbour required aid, and since I did not know this, it is incoherent to say that I had a reason to act.

This sort of response would be fundamentally misguided, however, since the equation between “X has a reason to φ” and “X could φ” is not warranted, given what is meant by X’s “having” a reason in this context. The mistake here rests on conflating this sense of having a reason with a second sense, a sense which refers to the extent to which some particular set of facts are epistemically accessible to, or actually known by, the relevant moral agents. It is this second sense, not the first, which carries implications for moral blameworthiness, since we must evaluate particular agents within their particular contexts- I may have a reason to φ in the first sense, without being morally blameworthy for failing to φ, if it is the case that I did not have access to sufficient facts to properly motivate my actions (and if it is the case that I am not to blame for this lack of access, perhaps because there was nothing I could do to acquire the relevant information for example).

Mason’s distinction between two kinds of ‘ought’ seems to map (at least in a rough sense) onto the above distinction between an objective and a subjective perspective about moral reasons. Mason uses the example of a woman who promises to visit a friend, but is prevented from doing so due to factors beyond her control (Mason, 2004, p.257). The woman still has a reason to visit her friend in the first sense; in the ‘abstract use of ought’ we might say that she ought to have visited her friend, without implying anything one way or the other in terms of a judgment about what she actually did. What we mean is that she had what Mason calls a “moral reason” (Mason, 2004, p.257); that there are facts we can point to which explain why, if circumstances had not prevented her, the woman ought to have visited her friend. Without
a clear distinction between these two senses of what it means to have a reason to do something, it’s plausible to imagine someone equating “this is what the woman had a reason to do” with “this is what the woman ought to have done”, and finding the result incoherent, given that there was no way for her to visit her friend.

What I want to emphasize in making this comparison between the two perspectives on reasons and the debate over the nature of justice, is that some confusion and disagreement over talk about moral reasons can be resolved by showing that the apparent incoherency stems from a conflation of two different ways of thinking about what it means to say that a person has a reason to do something. It turns out, once we are very clear about precisely what it is we are saying, that neither approach is in conflict with the other; we can talk about moral reasons in an abstract, objective sense, or we can take a subjective, perspective, and there need be no incoherency or incompatibility between the two provided that we are clear about the perspective we are taking. 20

If I am able to show that A-Type approaches are in fact coherent, I believe this comparison will help to explain why A-Type theories might seem incoherent to many of their critics. The idea here is that thinking that claims about moral reasons must be sensitive to the epistemic constraints that bear upon particular moral agents may be analogous to thinking that claims about justice must be sensitive to the various constraints that bear upon particular moral agents. If I can show, as I will attempt to do below, that we can ‘detach’ two different ways of talking about justice in a way which is similar to the two different perspectives we can take when talking about moral reasons, this may go a long way to explaining why the charge of incoherency against A-Type approaches is mistaken.

So, why should we think that the axiological approach can be a coherent one? Consider first, some basic ethical theory resembling a form of two-level utilitarianism along the lines developed by R.M. Hare. 21 This sort of approach might be appealing to us if we believed that a more straightforward brand of act utilitarianism is infeasible in practice. Say we believe that the point of ethics is to maximize the good and so we begin by affirming the truth of a form of act utilitarianism, but it turns out that our attempts to implement this theory are frustrated

---

20 It should be emphasized that my use of “objective” and “subjective” here apply to the perspectives we take when talking about reasons, rather than to metaethical claims about objectivity or subjectivity of reasons themselves. Here, to take the objective perspective means to report facts about the existence of moral reasons (regardless of how we describe their metaethical status), and to take the subjective perspective means to report facts about the epistemic accessibility of reasons to the relevant moral agents.

because, for example, human beings tend to lack the kinds of cognitive capacities that are necessary and sufficient to determine which actions will produce the maximum amount of good. Suppose we believe these cognitive incapacities to be fundamentally insurmountable (making these hard constraints, in this case).

In response to these feasibility constraints, we could decide to apply an act utilitarian approach to the question “are we likely to produce better overall consequences if we continue to try to act only such that we maximize the good?” We could then decide that we are likely to achieve the best consequences overall, if we rely on something other than a strictly act utilitarian approach. Suppose, after further analysis, we decide that we are most likely to achieve the best consequences overall if we behave exactly as some particular sort of Kantian would behave when it comes to ethics. The recommendations that our theory will now produce will be indistinguishable in practice from those that would be produced by the Kantian theory we are mimicking.

Thus far, this account might seem to give weight to Farrelly’s view, since the “correct” ethical position has been reached by responding to real world feasibility constraints. It turns out that to fully answer the question “how should we act towards other people?” we had to consider facts about human cognitive capacities and shape our theory accordingly.

Suppose, however, that both our pseudo-Kantian and a ‘real’ Kantian each affirm that we ought always to keep our promises, for example. It seems clear that while they will both issue the same recommendations, they will do so for quite different reasons; they will agree about what we must do, but not about why we must do it. My claim is that the discussion that follows between the two theorists will involve A-Type theorizing, and that such a discussion appears to be straightforwardly coherent. More generally, when philosophers disagree about whether consequentialist approaches are better than non-consequentialist ones, or virtue-based approaches, etc., there is usually no significant disagreement about the empirical, non-normative facts that apply in the real world. Rather, the disagreements tend to be about which facts are normatively significant or more significant than others, which principles are worth affirming and in what form, and so on. When we consider whether or not we ought to push the fat man off the bridge, in the classic trolley case, feasibility considerations generally don’t play a role in our deliberations- it is equally feasible for us to respond to consequentialist or deontic reasons (or others) in such cases, and the question of which reasons are relevant and
which are not will involve theorizing in a way which looks obviously coherent (I develop this point further below).

Farrelly could accept that there is indeed a distinction between arguments about how we ought to implement a value, and arguments about what sorts of values we should affirm in the first place, while insisting that ‘justice’ is the proper label for questions of the former sort but not the latter. This would be an extraordinarily weak response, however, since the dispute then becomes a mostly trivial one about what sort of labels we should assign to different sorts of project. Provided that it is accepted that there are two different sorts of project going on, the heart of the dispute is resolved. Beyond that, it hardly matters what particular labels we assign to each sort.

In addition to the example of cases where theories recommend identical actions for different reasons, another reason to think that the Axiological approach can be coherent is to consider the coherency of approaches to normative problems that rely on counterfactual reasoning in general. It seems clear that we can ask coherent questions which explicitly assume that the world is other than it really is, and produce useful results, without issuing any recommendations, except the sorts of conceptual recommendations I noted earlier. Take, for example, Judith Jarvis Thomson’s famous Violinist thought experiment. Briefly, you are to imagine that you awake in a hospital bed to find yourself connected to a famous violinist, who needs to remain attached to you for nine months, in order to be healed of an illness. The aim of this sort of argument is not to issue recommendations for how people ought to act should they ever find themselves hooked up to a violinist without their consent, or even to prescribe particular laws with relation to abortion more generally. Rather, the only recommendation that necessarily follows, if one is convinced by the argument, is that people should reject the principle that the woman can never override the right to life of the foetus if the foetus is considered to be a person. Whatever about the merits of this specific thought experiment, the approach in general seems an obviously coherent one.

It might be objected that these sorts of thought experiments are not really counterexamples to Farrelly’s claim, because they remain questions embedded within particular contexts. The fact that those contexts are different from the world we find ourselves in, so the objection might go, does not rebut the claim that these questions must be embedded in some context. In

22 This is true even if one thinks that the more extravagant our thought experiments, the less useful they are. See, for example Wilkes, (1988) and Brown & Fehige (2014).
fact, the only reason why we think that conclusions from thought experiments bear any relevance for us in our world depends upon the extent to which we can create a conceptual ‘bridge’ from the world the thought experiment takes place in to our own. By this, I mean that the worlds in which thought experiments take place must be sufficiently like our own, in order for their conclusions to affect what we should do or think in relation to our world and the people within it.

In the context of axiological approaches, however, this sort of objection does not suffice, for two reasons. Firstly, while I believe it is correct to say that thought experiments must be sufficiently ‘translatable’ to our own contexts in order to be useful to us, they are not necessarily rendered incoherent if we are unable to do so. We can coherently imagine what justice might be like for beings very different to ourselves; it’s just that this is not likely to be at all useful to us in figuring out what we ought to do or think when it comes to the real world. Clearly this is not a very strong response to the objection, since a coherent but useless approach does not seem like a particularly valuable one.

A much stronger reason to resist the objection is to emphasize, again, that Axiological approaches are concerned with which values and principles we should affirm and why, and the nature of and relationships between the values and principles we affirm. Given this, we can accept the requirement made by the objection which holds that thought experiments must be embedded within a context that is sufficiently similar to our own. In the case of Axiological approaches, however, it will turn out that the bar is set extremely low, in terms of what counts as a sufficiently similar context. The limitations are essentially only those that also determine the bounds of our own conceptual capacities- if we want to know what we should think about values, we cannot recommend that we should think something that it is impossible for us to think. This represents the hard constraints to do with the laws of logic, which were acknowledged as being present in A-Type theory and is merely a concession that we cannot imagine what we cannot imagine.

In other words, since A- Type theories are only concerned with telling us what we should think, provided that the A- Type theorist does not ask us to imagine something that is impossible to imagine, this ‘sufficiently similar context’ condition will be satisfied. In contrast, a Practical/P-Type theory will need to take additional facts into account (these will likely be a range of hard and soft constraints), since it cannot recommend that we act in ways that are impossible. Both approaches must concede that ought implies can, but only P-Type
approaches are necessarily constrained by limitations on what we can do, whereas A-Type approaches are constrained only by limitations on what we can think.

Justice and A-Type Theorizing

It is worth pausing for a moment to deal with a particularly important objection that might be levelled against Cohen (and anyone who thinks that we can do A-Type theorizing about a value like justice). An objector may concede that A-Type theory is usually coherent and maybe even useful and necessary, but deny that this is so when the concept in question is justice. Justice, so the objection goes, is special in the sense that it is the value which emerges via the trade-off of other significant values (like freedom, equality, and fairness, for instance). So, let me say something in this section about how I think Cohen (and the rest of us) can understand justice as a distinct value.

A defender of Farrelly’s, or a similar view, may make the following claims: A-Type theorising can be perfectly coherent and useful when applied to many values, but we should not think that we can do A-Type theorising about the concept of justice in particular. This is because there is something intrinsic to justice as a concept, such that it necessarily involves issuing recommendations about what a person is owed, given the presence of non-ideal constraints, including the need to make trade-offs between other values. The right way to understand the concept of justice is precisely in terms of the principles that govern such trade-offs. Justice, therefore, does not stand apart from these other values as an intelligible value in its own right. It is only when we are faced with decisions about how to strike the right balance between other values, that the value of justice ‘emerges’. In other words, “justice” is what we get when we strike the right balance. Any attempt to describe the concept of justice as an independent value which can itself be weighed against these other values will either (1) be incoherent or implausible or (2) will merely amount to labelling some other, more familiar value, as ‘justice’.

The above presents a challenge for anyone who wants to defend the idea of A-Type theorising about justice. First, we need a description of the concept of justice that is not reducible to some other basic value, like equality, or fairness. Call this the irreducibility requirement. Additionally, we need an account of the concept of justice that is not merely
irreducible to other values, but which describes what justice is in a way that seems intuitively plausible.\(^{24}\)

First, I want to present an argument against the view that justice is what emerges once we make the right trade-offs between other important values. If I am successful in this, then whatever “justice” is, it can be an appropriate object of A-Type theorising, (though I will not yet have shown whether “justice” is reducible to some other value).

Let us say that there is a concept, call it Concept X, that consists of the principles that determine what we ought to do, given all relevant hard and soft constraints and a plurality of values, and the need to strike the right balance between these values when it comes to how we treat people. What I have in mind here is the concept that Rawls would call ‘justice’ but which Cohen would call ‘rules of regulation’.\(^{25}\) If Concept X is justice in the Rawlsian sense, then it cannot make sense to ask what justice would require in a case where no balance needs to be struck. This would be a case where there are no relevant constraints of any sort, either hard or soft, and where it is possible to satisfy all of our values without needing to make any trade-offs whatsoever. We might think of such scenarios as constituting “unrealistic utopias” (in contrast to the Rawlsian idea of a ‘realistic utopia’\(^{26}\) which I imagine as being a world where only hard constraints apply).

Would it be incoherent to ask what justice would demand in an unrealistic utopia? If not, then it looks as though we can make sense of the concept of justice in an A-Type context. However, I can think of two reasons why one might be inclined to say that it does not make sense to talk about justice in unrealistic utopias.

First, one might think that we cannot talk about what justice requires in the absence of the ‘circumstances of justice’ (Rawls, 1999a, pp.109-112). It is only once we introduce scarcity of resources, limited altruism and so on, that questions of justice arise, and none of these conditions holds in an unrealistic utopia.

Responding to this objection requires us to distinguish two different ways in which the circumstances of justice can fail to apply. First, the circumstances of justice can fail to apply because there are too many constraints upon us. In these cases, questions of justice don’t arise

---

\(^{24}\) Presumably, there are all sorts of ways we could describe justice in order to make it an irreducible value, but most of these would simply not seem to capture what we mean when we describe something as a demand of justice.


\(^{26}\) See Rawls (1999b) p. 7.
because, for example, we simply don’t have any resources to distribute, or because we don’t have access to political institutions, and so on. One way to understand the circumstances of justice is as describing a minimum threshold, beneath which questions of distributive justice never arise, simply because there is no way to act justly towards one another.

Another way in which the circumstances of justice can fail to apply, however, is when we have even fewer constraints than those which are assumed by the circumstances of justice: if we have unlimited resources or unlimited altruism, for instance. It is easy to see why this way of thinking about the circumstances of justice is relatively uninteresting to anyone engaged with P-Type theorising – if we want to produce recommendations we might be able to act upon, then why bother imagining other kinds of worlds that could never exist? Nevertheless, it looks as though it would be perfectly coherent (if not especially useful) to ask what justice would require in such cases. Infinitely altruistic people with unlimited resources at their disposal would presumably still find themselves presented with a range of options as to how to distribute those resources. In order to know how they ought to distribute, we would need to figure out what justice demands, even in an unrealistic utopia.

A second, and in my view more powerful, objection would hold that my above response is in fact conflating the requirements of justice with the requirements of morality more generally. It isn’t “justice” which would demand, for example, that everyone gets an equal share of the unlimited resources in an unrealistic utopia, but (some particular account of) morality more generally which demands it. In order to respond to this objection, I need to offer an account of justice which is (1) narrower than morality more generally and (2) not reducible to other values. In other words, the concept of justice needs to describe something unique; something that is not the whole of morality itself, nor merely another way of describing some more familiar morally significant value.

Cohen only ever offers a provisional definition of the concept of justice in *Rescuing Justice and Equality*, where he says that justice is “giving each person her due” (Cohen, 2008, p. 7). In his distinction between principles of justice and rules of regulation, Cohen certainly seems to treat justice as what we might call a ‘basic’ value, to be traded off against other values in determining the rules of regulation. Is “giving each person her due” a plausible description of a value which is unique and not reducible to other values, like fairness and equality? It certainly seems as though these values can be pulled apart, so to speak. Consider, for example, treating a person fairly because we value fairness for its own
sake, and treating a person fairly because we believe that that is what they are ‘due’. It looks as though we can at least be committed to the idea that giving people their due is valuable, while being entirely agnostic as to whether people are due equal or fair treatment.

However, even if one is not convinced that “giving each person her due” is a distinct value, it may nevertheless be irreducible to other values. This would be so if we imagine the concept of justice as referring to a particular subdomain of morality, such that ‘justice’ is about picking out a particular set of values (those we intuitively take to be relevant to questions of justice) and describing how those values interact in order to determine what a person is ‘due’. Crucially, this process can tell us what a person is due in the absence of any other constraints, prior to any trade-offs with other values, and in a way which makes justice supervenient on, but not reducible to, the values of which it is a composite.

In Cohen’s case, thinking about justice in either of these ways (as either a basic value or an irreducible composite value) may help explain the apparent tension between his luck egalitarianism and his advocacy for an egalitarian ethos. If justice is about “giving each person her due” then we might say both that each person is due compensation for the bad effects of brute luck, and that each person is due a particular kind of egalitarian treatment, represented by Cohen’s idea of an egalitarian ethos, when it comes to productive choices. If both of these positions can be derived from the more basic thought that justice requires giving each person her due, this could explain why Cohen advocates both luck egalitarianism and the egalitarian ethos, despite the latter not being entailed by the former.

Does this view of justice risk over-inclusiveness? We might be tempted to read “giving each person her due” as a description of morality in general, rather than justice in particular. One response might simply be to bite the bullet here, and widen the scope of principles of justice such that they would apply to virtually every aspect of our lives. Another response might be to expand on Cohen’s definition and say that justice is about “giving each person what she is owed by the state”, as opposed to morality more generally, which is about “giving each person what she is owed by other people”. This second response seems to capture the right description of the relationship between justice and morality – just as people may be under moral obligations to establish states, and states may then incur special duties of justice towards people. Thinking of justice as “giving each person what she is owed by the state”

---

27 I am grateful to Alan Hamlin for suggesting this this point.
28 See also Tomlin (2010)
makes justice a distinct subdomain of morality (rather than expanding it to the point of meaningless, as with the first potential response).\textsuperscript{29}

A-Type Approaches: Not merely Coherent, but Necessary?

One form of argument that most of us have mastered by the age of five or six, consists in asking the question “But why?” repeatedly, until one’s opponent (usually one’s parent) gives up in frustration. A much more sophisticated version of this argument can be found in G.A. Cohen’s “Facts and Principles”\textsuperscript{30} where he argues that any fact-sensitive principle must ultimately be derived from some deeper normative principle that is not itself dependent upon any fact. Cohen’s central argument goes as follows (paraphrased):

1. Whenever a fact F confers support on a principle P, there is an explanation why F supports P, i.e. an explanation of how F represents a reason to endorse P.
2. This explanation invokes or implies a more ultimate principle, which would be true even if F were false.
3. If we investigate the full “chain” of explanations, we must eventually arrive at a principle that is not itself supported by any further fact.\textsuperscript{31}

In Cohen’s own words:

…we may ask anyone who affirms a principle on the basis of a fact what further and more ultimate principle explains why that fact grounds that principle and, once that more ultimate principle has been stated, whether it, in turn, is based on any fact, and so on, reiteratively, as many times as may be required until she comes to rest with a principle that reflects no fact… (Cohen, 2003, p.218)

In the context of the dispute between the context-sensitive and context-free views of justice, the initial argument could be made that fact-sensitive principles must depend on fact-insensitive ones, that the latter are conceptually prior to the former and must therefore be coherent, if they are to form the basis for fact-sensitive principles.

\textsuperscript{29} One might object to this response by arguing that it renders a concept like international justice incoherent, and that seems implausible. My intuition is that while the response I suggest above may require a redefinition of precisely which parts of morality properly belong to international justice and which belong to morality in general, we could still identify a domain of justice which is concerned specifically with the moral obligations of states as international agents.


\textsuperscript{31} See Cohen, 2003, pp. 217-218
It may be objected here that Cohen’s point doesn’t necessarily lend weight to the idea that we can coherently theorize about justice without reference to the real world, because Cohen’s claims only apply to the consideration of single principles or values, and not to questions about how to balance or make trade-offs between multiple competing values. If justice is about what we must do, so the objection will go, then it looks highly likely that we will be faced with problems involving multiple values, and our theory must be sensitive to the particular context in which we find ourselves.

There are at least two responses we might make to this sort of objection. First, if we accept that justice is about what we must do, we need not accept that that is all that justice is about: justice is also about why we must do it. Indeed, there may be many cases where we are unable to answer the former sorts of questions before answering the latter. Consider the following case, for example:

Suppose we want to know how to distribute some bundle of resources and we are presented with two options, both equally feasible in the sense that we will not be prevented from fully implementing either option if we choose to do so. Option A will distribute the resources in such a way as to cause extreme unwanted physical suffering to a small group of people, but will also cause a large amount of sadistic pleasure to a second much larger group of people. Option B will distribute the resources in such a way as to cause extreme and wanted pleasure to a small group of people, but will also cause a large amount of anguish to a second small group of misanthropists who loathe the sight of others’ happiness.

In this case, since we have stipulated that feasibility is not an issue, determining which course of action to take will require an investigation as to whether or not there are any moral constraints that bear upon us, such as demands of justice or morality that would prohibit us from choosing one option over the other. In response to the above case, we could propose a principle that states that what matters is that we maximize the total amount of pleasure, which would lead us to choose Option A, since the pleasure of the sadists outweighs the pain of the small group. A competing (and more intuitively plausible) principle might hold that it is impermissible to cause suffering to one group of people, merely for the pleasure of others, or to withhold pleasure from one group of people merely to avoid causing misanthropists to suffer, in which case we would be bound to choose Option B.

In order to determine what we ought to do in this case, the axiological approach is required, and there are likely to be a great many real-world cases where a similar situation
applies. This is because, even in the real world with its ubiquitous feasibility constraints, it is almost never the case that we have only one viable option available to us. In fact, if only one course of action is feasible, the question as to what justice demands seems no longer relevant—just as if ought implies can, and there is only one thing we can do, then there is probably not much worth debating at that point. In most cases, however, where we are presented with multiple options, even if we impose harsh feasibility restrictions, we may still be left with the task of choosing between two or more equally feasible courses of action. At this point, it remains coherent to ask what principles we ought to endorse, given the available options, and without further consideration of feasibility constraints.

It might be objected in response to this that the above sorts of cases do not threaten Farrelly’s view, given that we are deciding between competing principles only after having first limited our options to those principles which could be feasibly implemented. So, Farrelly could simply say something like “of course, that is a coherent sort of discussion to have, but only because we have already taken feasibility considerations into account. Once feasibility is factored into the equation, we have a necessary but not sufficient component of justice. It would be silly to deny that further theorizing is required, given that we are not interested merely in what we could do but what we should do.”

This sort of objection is almost persuasive, except that there is no reason for us to think that the coherency of asking what principles we should affirm need come after we have taken feasibility considerations into account, rather than before. If the claim is merely that feasibility considerations must play some role before we arrive at action-guiding recommendations, this is something that is not merely compatible with but explicitly acknowledged by both Rawls’ and Cohen’s arguments. As has been noted earlier, Rawls envisions his version of ideal theory as the first of a two-stage process, motivated explicitly by real world problems. Cohen’s distinction between the principles we should affirm and the rules of regulation which attempt to implement them in the real world also renders his view compatible with the above sort of objection.

The idea that the axiological approach may be a necessary conceptual antecedent to fact-sensitive recommendations is compatible with acknowledging the various difficulties that come with trying to balance and make trade-offs between different principles and values and

---

with acknowledging that Cohen’s arguments only apply to single values or principles. Regardless of these difficulties, it is coherent to ask whether some particular principle or value ought to be part of our considerations at all. So, for example, we might believe that a value like “liberty” must be considered alongside a value like “welfare”, but we can coherently put forward arguments as to why these values should form part of the equation, rather than values like “promoting suffering and misery” or “promoting dishonesty” and so on. The axiological approach is necessary to perform this function, even if we remain uncertain as to how to make trade-offs between competing values. The A-Type theorist can concede that feasibility will be important when it comes to devising what Cohen calls “rules of regulation”34 or when it comes to implementing values in the real world, but there are good reasons to think that the feasibility-sensitive aspects do not exhaust justice as a concept. It appears that the axiological approach has a key role to play, both in its ability to identify which principles are important at the initial, abstract stage of the theory, and when it comes to deciding between equally feasible options in the real world, once additional feasibility constraints have been taken into account (since the reasons we have for choosing one option over the other in such cases will not be “feasibility-sensitive”).

Additionally, the coherency of the axiological approach can be further demonstrated by confining it to the investigation of a single principle; the principle that feasibility is important when it comes to theorizing about justice. If we ask why it’s important that we care about feasibility, there are two sorts of answers we could give. The first sort will point to concerns about injustice in the real world; we must care about feasibility because we must care about injustice in the real world, and if we don’t care about feasibility we will be much less likely to address injustice in the right sort of way. That we should care about injustice in the real world seems to be a normative claim which results from an axiological approach which is not itself dependent upon concerns about feasibility; we care about the feasibility of our theories only (or primarily) because we care about injustice.

2.3 Are A-Type Approaches Valuable?

If A-Type approaches are not merely coherent but also necessary in some contexts, then this gives us obvious reasons to value A-Type theorizing. Clarifying the relationship between

ideal and non-ideal theory, as Farrelly attempts to do, involves engaging in the sort of conceptual clarification that I have argued belongs to the domain of A-Type theory (or the theory of ideals). I also argued that A-Type theory will be necessary in cases where we must choose between equally viable options that require rejection or affirmation of particular normative principles, depending on which option we choose.

Furthermore, it seems clear that sometimes there can a threshold below which we would know so little about our values and principles, that we would have no non-arbitrary basis upon which to identify the sorts of problems we want to investigate let alone the sorts of actions we ought to take to implement them. It could be the case that any attempt to investigate our values and principles from a point of such profound initial ignorance would be doomed to cause more harm than good, but we would have no way of knowing so. Inaction in such a scenario would be indicative of a kind of action motivated by some normative beliefs, whether we realized it or not. So, it cannot be plausibly suggested that achieving a clearer understanding of our values and principles is pointless in the sense that it never serves a useful or necessary purpose.

Are A-Type Approaches no longer necessary? The Resource Allocation Objection

One objection which remains open to the critic of overly-abstract approaches (both A-Type and certain kinds of P-Type approaches such as those of Rawls and Dworkin, for example) would be to say something like the following: “Additional A-Type theorizing is pointless in the context of political philosophy, because we are sufficiently clear about our values and principles such that we have all the intellectual and conceptual resources we need to tackle the most urgent problems which present themselves in our non-ideal world.” One problem with this claim is that there are ongoing fundamental disagreements about what our values and principles should be and the relationship between different values and principles that, depending upon the position one adopts, are likely to have implications when it comes to acting upon those values and principles in the real world.

In order to give the objection a fairer chance, therefore, we must offer a weaker claim. Rather than claiming that we are sufficiently clear about our principles, let the claim be instead that there is a disproportionate emphasis upon abstract issues in political philosophy,
relative to the attention given to solving real-world, non-ideal problems, especially problems where there is a consensus amongst theorists as to the values we ought to affirm and implement. Human slavery, for example, is probably not the sort of issue that requires further theorizing at a very abstract level in order to guide our actions. In places where such obviously unjust practices persist, however, it would be strange to think that this is typically due to the failure of political philosophers to apply themselves to non-ideal solutions to the problem. On the other hand, if political philosophy is to be useful in an action-guiding sense, then there must be less obvious issues of injustice that require philosophers to illuminate the issues, clarify the questions at stake, and communicate the answers to these questions to the wider community in order that the injustice is eliminated. The refined version of this sort of objection then, which I shall call the ‘Resource Allocation Objection’, holds not that A-Type theorizing is pointless or unnecessary, but that many or most political philosophers are misusing their intellectual resources by failing to pay sufficient attention to those cases where progress can be made here and now in the real world. If this is the sort of objection that Farrelly and other critics of ideal theory have in mind, then it is important to distinguish this kind of objection from those objections that suggest a deeper, more fundamental problem with the coherency and/or usefulness of highly abstract theorizing in general.

This sort of objection may serve as the chief motivational force behind many of the attacks on A-Type approaches. Distinguishing claims about the coherency and usefulness of A-Type theorizing from claims about its pervasiveness, allows us to reject objections against the former while leaving room for the possibility that there may be nothing wrong with A-Type theorizing except that there’s simply too much of it. I will not offer an argument against this form of the resource allocation objection here, partly because it does not touch on the deeper questions about the usefulness of A-Type approaches in general, and partly because I suspect it may simply be correct, at least to some extent.

Are A-Type Approaches ‘Dangerous’?

One way in which A-Type theorizing might be not merely unhelpful but actually counterproductive to the pursuit of justice in the real world would be if the A-Type theory attempted to produce recommendations for the real world, without taking hard and soft constraints into account. So, for example, we might expect a good theory of justice to begin
with an account of the principles and values we should affirm (A-Type theory) before moving towards a Rawlsian-style account of what justice would look like in a realistic utopia (P-Type theory with hard constraints only), before moving towards the fully non-ideal real world by taking more and more soft constraints into account (P-Type theory with hard and soft constraints). In contrast, we can imagine a theory that begins with the A-Type approach and issues recommendations about how we ought to act in the real world, without taking either hard or soft constraints into account. The dangers of such an approach are obvious; we might end up with recommendations that we cannot possibly implement and/or recommendations that in fact have the effect of making the world a more unjust place.

It’s not plausible, however, to suggest that any A-Type theory must end up producing these sorts of misguided recommendations instead of, or in addition to, what it is normally supposed to produce (i.e. conceptual clarity about our values and principles, an account of the reasons we have for affirming them and the relationship between them). Why might some A-Type theorists make this mistake? According to the constraint-based view, if the theory has A-Type purposes, if the correct inputs are present in the theory, and if there are no errors in reasoning along the way, the outputs which are produced will be recommendations about what we should think, rather than recommendations about what we should do. This suggests that an A-Type theory can fail if there are errors in reasoning and/or if there are mistakes in the assumptions from which the theory begins (either because the theory assumes things that it should not, or because it fails to make assumptions that it should).

The objection that an A-Type approach will be counterproductive or dangerous if it produces infeasible recommendations should therefore be seen as opposing bad A-Type theory rather than A-Type theory in general. A-Type theorists can minimize the chances of such mistakes by being as clear as possible about the purpose behind the theory, about the sorts of assumptions that are being included and excluded as part of the inputs, and by offering a clear and well-reasoned set of arguments leading from the inputs to the outputs.

One upshot of this call for clarity is that it actually undermines the argument that we shouldn’t pay as much attention to A-Type theories in general, even in cases where we believe our understanding of our values and principles is sufficient to allow us to engage in non-ideal problem-solving. Consider that there may be cases where we may believe that we have sufficient understanding to proceed only to discover upon further reflection that we were mistaken, in which case our actions in the real world may have been futile or even made
the situation worse. A-Type theories can be damaging if they produce recommendations that we ought not to act upon directly. However, they may also be damaging when they produce conclusions about our values and principles that turn out to be false. It seems prudent, therefore, to ensure that we continue to maintain some level of scrutiny when it comes to A-Type theorizing, so that we can increase our chances of spotting a faulty theory whether it gives us bad recommendations or bad information about our values and principles. This will require us to continue to engage with A-Type theories even in cases where these theories seem to provide us with a sufficient foundation from which to build towards solving real-world problems.

Recall that in the previous chapter the Critics’ Argument was presented as follows:

P1 – We should reject any method of theorizing about justice that fails to, or is not concerned with, providing us with action-guiding principles in the real world.

P2 – The Rawlsian method (or any variant that relies heavily upon ideal theory) fails to provide us with action-guiding principles in the real world.

C – We should reject the Rawlsian method (or any similar variant that relies heavily upon ideal theory).

We can now conclude that the first premise above is false. We should not reject any method of theorizing about justice that is not concerned with providing us with action-guiding principles in the real world that are both achievable and desirable. We should not reject such approaches because there are areas of justice where we still lack sufficient understanding about what our values and principles ought to be, and because it is prudent to ensure that we continue to subject those values and principles we do affirm (even those about which there is virtually no reasonable disagreement) to scrutiny. We also ought to ensure that when we engage in theorizing of any sort that we are as explicit as reasonably possible about the sorts of inputs we are using, how we are using those inputs to produce our outputs, and what sort of outputs our theory is intended to produce. We might also wish to see political philosophers distribute their intellectual resources in some different way if it is the case that more good would be achieved by paying more attention to certain non-ideal problems. Yet these recommendations do not amount to the rejection of A-Type theorizing called for in the first premise of the Critics’ Argument; at most, they support reform rather than rejection.
2.4 Why Cohen is an A-Type Theorist

It’s important to re-emphasize at this point that Farrelly either does not accept the A-Type/P-Type distinction, or if he does, he does not accept that Cohen’s theory should be categorized as the former rather than as a (relatively extreme) form of the latter. Farrelly primarily focuses his attack on ideal theory against what he calls the “moderate idealized position”. He explains his choice of target in a footnote where he says “I have chosen the moderate idealized position as my target, rather than the extreme idealized position (e.g. Cohen), as I believe that the case against the moderate position is much harder to make.” (Farrelly, 2007, p.862) Obviously, this implies that Farrelly thinks the case against Cohen (and other A-Type theorists) is relatively easy to make.

I hope to have demonstrated in the previous sections of this chapter, that this is not true if one accepts the A-Type/P-Type distinction, and provided that one agrees that Cohen’s approach is an example of the former. What Farrelly means by “the case against” Cohen, therefore, is most plausibly understood as the case against Cohen as a P-Type theorist, as a position lying on the extreme end of a spectrum with theorists such as Rawls in the middle. This is why Farrelly targets Rawls rather than Cohen- if Cohen is merely a more extreme version of Rawls, and if Farrelly can successful make the case against Rawls, then Cohen’s position is automatically ruled out, provided that it depends upon the same foundations as that of Rawls. (Farrelly says “if one can show that even moderate ideal theorizing is inherently flawed then one has also shown that the extreme idealized position is flawed.” ((Farrelly, 2007, p.862)))

So, in one sense, none of the previous section serves as a direct response to Farrelly, since he implicitly rejects (or fails to consider) the characterization of Cohen’s approach as A-Type theory. However, the previous sections serve at least one useful function in constructing the case against Farrelly; if my defence of A-Type theorizing in the previous section seems plausible, then it removes the option for Farrelly of simply accepting the A-Type/P-Type distinction, accepting the positioning of Cohen’s approach as A-Type theory, and then launching a new set of arguments against Cohen as an A-Type theorist.

The first part of this chapter aimed to establish that the distinction between A-Type and P-Type theories is coherent, and the second offered reasons to think that A-Type theorizing is
worthwhile. My final task in this chapter is to demonstrate that we should think of Cohen’s project as being an example of A-Type theorizing.

By ‘Cohen’s project’, I mean the sorts of activities Cohen is engaged in in his book *Rescuing Justice and Equality*, and in particular the chapter ‘Facts and Principles’ (which was previously published as a paper in its own right). These are the only works of Cohen’s which Farrelly cites in support of his characterization of Cohen as an ‘extreme’ ideal theorist. Of course, Cohen is free to engage in non-A-Type theorizing elsewhere; my aim here is to show only that *Rescuing Justice and Equality* and Facts and Principles in particular are examples of A-Type projects.

In ‘Facts and Principles’ Cohen states his thesis as follows: “a [normative] principle can reflect or respond to a fact only because it is also a response to a [normative] principle that is not a response to a fact” (Cohen, 2008, p.232). Cohen defines a normative principle as “a general directive that tells agents what (they ought, or ought not) to do” (Cohen, 2008, p.229). Facts, on Cohen’s view, are truths, other than normative truths, which someone might reasonably think supports a particular principle.

Cohen is at pains to point out that his thesis is neutral with regard to various metaethical questions. He claims his thesis does not depend upon a particular conception of what sort of thing a fact is, of whether or not an “ought” can follow from an “is”, or the objectivity of normative principles and questions about moral realism, for example.

The position Cohen takes himself to be arguing against is one he attributes to Rawls amongst others, which holds that what we believe about ultimate normative principles must depend to some extent upon non-normative facts about human nature, or as Cohen puts it “that all principles for governing human life are sensitive to facts about human life” (Cohen, 2008, p.232). Cohen sets out his argument against this position as follows:

P1 – Whenever a fact F confers support on a principle P, there is an explanation why F supports P, an explanation of how, that is, F represents a reason to endorse P.

---

35 Farrelly also cites Cohen’s “On the Currency of Egalitarian Justice”, as an example in his discussion of luck egalitarianism (he considers luck egalitarians to be working at the level of ideal theory), but as he does not engage with that paper directly it should be sufficient for me to do likewise.

36 Facts and Principles was originally published as a paper in 2003 and a revised version included as part of Cohen’s 2008 book *Rescuing Justice and Equality*. The references here are to the paper in its latter form.

37 Cohen’s claim that his view is metaethically neutral is controversial, however. See, for example, Ronzoni & Valentini (2008).
P2 – The explanation whose existence is affirmed by the first premise invokes or implies a more ultimate principle, commitment to which would survive denial of F, a more ultimate principle that explains why F supports P.

P3 – The sequence [of explanations] cannot proceed without end because our resources of conviction are finite, and even if they were not, proceeding without end would violate the self-understanding stipulation. [The self-understanding stipulation requires that the person we ask for an explanation has a clear grasp of what their principles are and why they hold them.]

Farrelly characterizes this position as the belief “that the fundamental principles of justice are logically independent of issues of feasibility and questions about human nature” (Farrelly, 2007, p.844). If we combine this with Cohen’s definition of a normative principle as “a general directive that tells agents what they ought to do” (Cohen, 2008, p.229), it might well seem that, protestations to the contrary notwithstanding, Cohen’s position must be potentially vulnerable to issuing incoherent directives which demand that we act in ways which are simply impossible.

This thought overstates the nature of Cohen’s thesis in “Facts and Principles”, however. Cohen stresses that his claim is strictly one about the logical structure of a coherent set of beliefs. It is in this sense descriptive. Farrelly’s worry is clearly that, by beginning to think about principles in the most abstract sense possible, we will arrive at prescriptions for action that are either impotent or counterproductive. But according to Cohen:

…the priority enjoyed by fact-insensitive principles is purely logical, and not temporal or epistemic, or at any rate not epistemic in at least one sense of that term. The priority of fact-insensitive principles is a matter of what utterances or principles commit one to, not of how one comes to believe or know what one says in uttering them. That is why I do not deny (in fact, I would assert) that asking what we think we should do, given these or those factual circumstances, is a fruitful way of determining what our principles are; and sometimes, moreover, responses to actual facts reveal our principles better than our responses to hypothesized facts do…But none of those considerations bear on whether commitment to fact-sensitive principles carries with it commitment to fact-insensitive principles (Cohen, 2008, p.247).

Cohen says explicitly in the above passage that asking what we think we should do, given certain factual circumstances, is a fruitful way of determining what our principles are. This seems to be quite similar to Farrelly’s insistence that action-guiding recommendations must be appropriately fact-sensitive. One possible source of confusion might lie in Cohen’s description of normative principles as directives that tell us what we ought to do. There may
be some ambiguity here, however, as to what it might mean for a principle to tell us what we ought to do.

Consider that Cohen characterizes a distinction between his approach to normative theory and that of Rawls and others by telling us that he views political philosophy as being primarily concerned with “what we should think, even when what we should think makes no practical difference,” rather than “what we must do” (Cohen, 2008, pp.268-269). There is, however, at least one sense in which “thinking” is a kind of “doing”. To place value upon a principle, or to affirm some value as more or less important as another, for example, is an act that can have a normative dimension. It is surely true that some normative principles can be concerned with the kinds of things we ought to value, rather than how we should implement those values in the real world.

By assuming that Cohen’s claims about the logical independence between fundamental normative principles and fact-sensitive questions about how best to implement those principles are vulnerable to issuing misguided recommendations, Farrelly seems to be assuming a temporal and epistemic priority of the kind Cohen explicitly rules out. Furthermore, he seems to be assuming that Cohen’s thesis in “Facts and Principles” is issuing a recommendation of its own; that we ought to begin by theorizing about justice (almost) without constraints, and then seek to implement our conclusions in the real world.

On the contrary, the recommendations which are produced by Cohen’s argument here do not tell us how we ought to go about addressing problems of injustice in the real world, or even about how we ought to go about articulating justice in an abstract sense. Rather, the implied recommendation is merely that we should think about every fact-sensitive normative claim as being ultimately dependent upon a normative claim that is fact-insensitive.

Given the above, we are now in a position to offer a firm answer to the question of whether or not Cohen ought to be regarded as an A-Type theorist. As I have defined it, there are two features that define the A-Type approach. The first is the absence of constraints as part of the inputs. Given that Farrelly himself characterizes Cohen’s approach as supporting the idea “that the fundamental principles of justice are logically independent of issues of feasibility and questions about human nature” (Farrelly, 2007, p.844), it is fair to say that Farrelly must accept that this characteristic of A-Type theorizing applies to Cohen’s approach.
The second characteristic is that A-Type theories seek to produce recommendations about how best to conceive of a concept (in this case, ‘justice’) rather than to produce action-guiding recommendations. The best evidence to support the claim that this is a characteristic of Cohen’s project is Cohen’s distinction between ‘principles of justice’ and ‘rules of regulation’ and Cohen’s distinction between his approach to political philosophy in general, and that of Rawls.

To begin with the latter, which will lead us to the former, in the introduction to *Rescuing Justice and Equality* under a section titled “Methodological Disagreements”, Cohen notes a disagreement between himself and Rawlsians about how to do political philosophy and (amongst other things) “the relation of theory to practice”. Cohen contrasts his preferred approach (“the method that I favour investigates the shape of, and, consequently, the logical implications of, our deepest normative convictions” ((Cohen, 2008, p.7))) with what he labels as Rawls’ “constructivist” view:

On the constructivist view, the content of justice is identified by the rules for social living, the rules of regulation that would be chosen in a privileged choosing situation…My rescue of justice denies the mooted identification on two grounds: if rules for social living are soundly based, they will reflect both values other than justice and practical constraints that restrict the extent to which justice can be applied. That being so, justice, itself, could not be what is specified by such rules (Cohen, 2008, p.3).

Cohen goes on to say:

More generally, the Rawlsian approach denatures justice, since it cannot recognize that if something is unfair, then it’s to that extent unjust: *the identification of the best-all-things-considered rules of regulation with principles of justice excludes that recognition* [emphasis added] (Cohen, 2008, p.8).

Cohen’s aim then, is to establish a distinction between principles of justice and rules of regulation, and to criticize Rawls on the grounds that he incorrectly conflates the two. Cohen lays this out most clearly when he says:

The Rawlsians, who believe that the constraints of human nature and human practice affect the content of justice, are inclined to regard me as unrealistic and/or utopian in that I believe that justice is unaffected by those mundanities. But it is worth pointing out that they are in one way more Utopian than I. For in believing that justice must be so crafted as to be bottom-line feasible, they believe that it is possible to achieve justice, and I am not so sanguine. It follows from position that justice is an unachievable (although nevertheless governing) ideal (Cohen, 2008, p.254).

Ironically, the above passage positions Rawls as being closer to Farrellly than Rawls is to Cohen. Following the above, the reason Cohen does not take constraints into account (the first of the two characteristics of A-Type theories) is because it is not merely sufficient but
necessary in order to achieve the kinds of outputs he desires, which concern an account of the
deal of justice, independent of the rules of regulation which might constrain the application
of that ideal in the real world.

2.5 Conclusion

Colin Farrelly’s arguments against ideal theory place theorists like Cohen and Mason at
the extreme end of a spectrum of approaches with theorists like Rawls and Dworkin
occupying a “moderately” ideal position (Farrelly, 2007, p.845). Farrelly correctly assumes
that if his arguments can defeat these moderate positions, they will simultaneously defeat the
more extreme versions as well. In this chapter, my aim has been to show that Farrelly errs in
conceiving of theorists like Cohen and Mason as belonging on the same spectrum as theorists
like Rawls and Dworkin.

While Farrelly acknowledges that Cohen and Rawls differ in terms of the kinds of
feasibility constraints they take into account when it comes to the inputs of their theories,
Farrelly fails to recognize the distinction between Cohen and Rawls in terms of the intended
purpose of their projects. To this end, I argued that there is an additional distinction between
A-Type theories and P-Type theories such that A-Type theories are intended to produce
conceptual clarity about normative ideas and are explicitly not concerned with producing
action-guiding recommendations about how those ideals should be implemented, in contrast
with P-Type approaches.

I noted that there are three lines of response open to Farrelly, given the above criticism.
First, Farrelly could deny the coherency of the distinction between A-Type and P-Type
theories. In the first section of this chapter, I defended the claim that the distinction is a
coherent one, and suggested that Farrelly fails to distinguish between two senses of a
“normative theory” and/or, following Mason, that he fails to distinguish between two senses
of “ought”; one which implies a judgment about the moral character of an individual, and one
which implies a claim about objective moral facts which are independent of considerations
about particular constrained agents. The thought that an analysis of an ideal that produces no
action-guiding recommendations is incoherent can be explained by a failure to make these
kinds of distinctions.
The second kind of response open to Farrelly would be to accept the distinction between A-Type and P-Type theory, but to deny that Cohen’s approach is an example of A-Type theorizing. In the third section of this paper, I aimed to show that Cohen’s approach clearly fits the definition of A-Type theorizing. Farrelly himself acknowledges that Cohen’s approach is not concerned with constraints as part of its inputs, and Cohen’s entire project in *Rescuing Justice and Equality* turns on Cohen’s distinction between his approach and that of Rawls. This distinction is between an approach that conflates justice with rules of regulations and Cohen’s preferred method of investigating the ideal of justice without feasibility constraints in order to produce conceptual clarity rather than action-guiding recommendations.

Finally, Farrelly could accept that the A-Type/P-Type distinction is coherent, and that Cohen’s project is an example of the former, but reject the idea that A-Type theorizing is worthwhile. In Section 2 of this chapter, I argued that A-Type theorizing has played and continues to play a useful and necessary role in thinking about normative subjects. I also argued that any danger caused by a mistaken attempt to directly apply A-Type theorizing to the real world can be avoided by theorists striving to be clear and explicit about the aims and structures of their approaches, and that this call for clarity does not amount to a good reason to abandon A-Type approaches in general.
3. The Constraint-Based View

3.1 Introduction

The aim of this chapter is to describe and defend my preferred conception of the relationship between ideal and non-ideal theory, which I shall call the ‘constraint-based view’, or ‘CBV’.

The CBV is a view about how to conceptualise the differences between different kinds of P-Type theories, i.e. theories that are part of a process intended to help us produce action-guiding recommendations that are achievable and desirable. Some critics of ideal theory, notably Colin Farrelly have tended to place A-Type and P-Type approaches in a single category of ideal theories, with A-Type theories being seen merely as a more idealised form of P-Type theories (Farrelly, 2007, p.846). The aim of the previous chapter was to insist upon a distinction between A-Type and P-Type theories, thereby allowing me to be clear about the scope of the view I will defend in this chapter.

In Chapter 1, I described several different ways of conceptualising the relationship between ideal and non-ideal theory, and noted problems with most of these views, with the exception of Hamlin and Stemplowska’s ‘multi-dimensional continuum’ view. In Section 3.2 of this chapter, I will describe the CBV, before returning to the alternative views in Section 3.3 of this chapter, in order to show the advantages of endorsing the CBV over these approaches.

38 Throughout this chapter, I will sometimes refer to both the “relationship” and the “distinction” between ideal and non-ideal theory. With the exception of Hamlin & Stemplowska’s view, most of the views with which I will be comparing the constraint-based view conceive of a fairly sharp distinction between ideal and non-ideal theory. However, on my view, it is more accurate to conceive of a spectrum of more ideal to less ideal approaches, depending on the kinds of constraints that are taken into account as inputs to the theory. Whenever I refer to ideal and non-ideal theory in a way which suggests a clear distinction between the two, I am either using the terms of these alternative approaches, for the purposes of an internal critique of those approaches, or I have in mind two ends of a spectrum – with extremely ideal approaches at one end, and extremely non-ideal approaches at the other. The idea of a spectrum of more and less ideal approaches, rather than a sharp distinction between ideal and non-ideal approaches is a key part of the constraint-based view that I describe in more detail in Section 1.
alternative views. The alternative views I shall consider are what I will call the Rawlsian view\textsuperscript{39}, Mills’ view, Farrelly’s view, and Sen’s view.

In some cases, the alternative view will be flawed because it is incapable of taking into account different dimensions of idealness or non-idealness. In this sense, it will be too narrow, compared to the CBV, which I will argue is broad enough to take into account everything we think is worth taking into account, in terms of the constraints that bear upon a P-Type theory. In other cases, the alternative view will turn out to be flawed for other reasons, which I shall explain as I go through the different views in Section 3.3.

3.2 – Describing the Constraint-Based View

Recall that one of the key distinctions between A-Type and P-Type approaches is that the latter are concerned with (eventually) producing action-guiding recommendations that are achievable and desirable. Since these intended outputs are recommendations about how real people ought to act in the real world, it seems correct to assume that such recommendations will be constrained by the maxim that ‘ought’ implies ‘can’\textsuperscript{40}. A recommendation that advises us to act in ways that are impossible is not a recommendation to do something that is achievable. One way to assess the worth of a P-Type theory, therefore, will be to assess whether it does, or is likely to, produce recommendations that can be acted upon.

The most obvious way of checking whether a P-Type theory passes this test will be to examine the inputs to that theory, or the assumptions that the theorist takes for granted\textsuperscript{41}. A theory that assumes that humans can do things that they cannot do will be more likely to recommend that humans do things that they cannot do. If that happens, we can say that the theorist has failed to take into account a relevant constraint upon human behaviour, and that

\textsuperscript{39}I call it “the Rawlsian view” rather than “Rawls’ view” here because it is not clear to me to what extent Rawls himself is necessarily committed to the Rawlsian view, but it does seem clear that this view has been endorsed by most philosophers who consider themselves Rawlsians.

\textsuperscript{40}While some people dispute that ought implies can, it seems clear that ought does at least sometimes imply can in certain contexts, and it seems reasonable to suggest that in the context of producing action-guiding recommendations for the real world, it is safe to assume that ought implies can when considering such recommendations.

\textsuperscript{41}Here I follow Stemplowska (2008) in conceiving of theories as consisting of a series of inputs, rules of derivation, and outputs.
this explains why her theory is defective in failing to produce the kinds of outputs it was intended to produce (i.e. action-guiding recommendations that are achievable and desirable).

A ‘relevant constraint’ in this sense is a type of fact that plays a role in explaining why it is impossible for some agent(s) to perform some particular recommendation. A ‘recommendation’ here just means some particular plan of action prescribed by the theory in question.

Some of these facts can be understood as ‘hard’ constraints upon human behaviour. They are constraints because they make certain plans of action impossible for human beings, and they are ‘hard’ constraints because we have good reason to believe that there is nothing we can ever do to alter those facts. Some hard constraints are more controversial than others, and a theory may be more or less conservative depending on what it views as being ‘up for grabs’. For example, the assumptions that human beings cannot travel backwards in time or be in two places at once are examples of uncontroversial claims about hard constraints. The assumption that human beings can never be motivated to care as much about distant strangers as they do their co-nationals, on the other hand, is an example of a highly controversial claim about a hard constraint.

Labelling a constraint as a hard constraint should be subject to a reasonableness criterion. There are some scientists who may believe that it is possible to travel at superluminal speeds, for instance, but it is not reasonable to offer this as an objection to a theory that assumes that it is a hard constraint that humans cannot travel faster than the speed of light. Similarly, a theory of justice that assumes that we are not brains in vats should not be considered faulty for that reason. If we are to do political philosophy at all, there are certain things we must be able to take for granted, so these kinds of unreasonable objections can be safely dismissed.

42 Throughout this chapter, I will use the phrases ‘hard constraints’ and ‘soft constraints’ to describe something very similar to the concepts described by Pablo Gilabert and Holly Lawford-Smith (2012) when writing on the topic of political feasibility. Unfortunately, I had chosen these terms before the publication of Gilabert and Smith’s paper, and I have not yet been able to find better terms to use to distinguish my view from theirs. Specifically, I believe our views differ in three important respects. Firstly, while our views are similar, the CBV is a view about ideal and non-ideal theory, while Gilabert and Lawford-Smith’s view is a view about political feasibility. This makes their view more general, since we can be interested in feasibility as a more general concept outside the context of ideal and non-ideal theory. Secondly, I include moral constraints in my taxonomy of constraints, while Gilabert and Lawford-Smith do not. Thirdly, I view every soft constraint as a potential hard constraint, given that most if not all constraints are not infinitely malleable. I return to Gilabert and Lawford-Smith’s view in more detail in Chapter 4, which focuses on the topic of feasibility.
So, if we want to produce a theory which does not recommend that people do the impossible, it will be necessary to identify and include all relevant hard constraints as inputs to that theory.

If a P-Type theory takes all and only relevant hard constraints as inputs to the theory, it lies at one end of a spectrum. We might call such a theory an example of a ‘pure’ ideal theory. In every relevant dimension, to use Hamlin and Stemplowska’s terms, the theory takes all relevant hard constraints into account, but only hard constraints.

At the other end of the spectrum is a theory that takes all relevant hard constraints into account, in addition to all relevant soft constraints. A soft constraint is a fact that explains why a particular course of action is impossible to achieve at present, but where it is possible to alter this fact, in contrast to unalterable hard constraints.43

By ‘taking a constraint into account’, I mean including the relevant fact as part of the inputs to a theory. When hard constraints are included, these serve to block any recommendations that would require that those hard constraints not be in place. Taking a soft constraint into account is somewhat different, however, given that soft constraints can potentially be weakened or eliminated. A more non-ideal approach will seek to identify the range of options open to us here and now, given those options permanently blocked by hard constraints and presently blocked by soft constraints. The ultimate recommendations that emerge from such an approach may well provide a plan of action aimed at weakening or eliminating some relevant soft constraint(s), in order to make possible more desirable plans of action.

Here are some simple examples of soft constraints: Suppose that the rates of breast cancer among women in our country are particularly high, and (motivated by some particular theory of justice) we want to come up with recommendations to reduce these rates. A soft constraint might be the fact that there is little public awareness about the issue and that breast cancer services tend to be underfunded. Responding to these soft constraints might produce recommendations to increase funding for breast cancer treatment and funding a public campaign instructing women how to check for lumps and informing them of how to avail of a free screening for breast cancer at their local hospital. Or suppose that there is a famine in

---

43 On this view, there is scope for disagreement between different P-Type views. For example, they can disagree about whether something is a hard or soft constraint, the strength of the constraint in question, or the normative assessments of the available options, given a similar account of all relevant hard and soft constraints.
some foreign country and (motivated by some particular theory of global justice) we want to provide relief to the victims. Suppose it turns out that neither our politicians nor the populace is sufficiently motivated to help the victims. A plan to tackle these soft constraints might include lobbying public officials and engaging in a publicity campaign designed to increase people’s empathy for the victims of the famine. Or suppose that (motivated by a desire to abide by the Difference Principle in our distribution of benefits and burdens in society) we need to know who the worst off people will be, following the implementation of some proposed new tax, but that we lack the empirical data necessary to identify the group that is likely to be worst off, and to understand just how badly off that group will be. In this case, tackling the relevant soft constraints might result in a recommendation to fund the relevant empirical research to provide us with the data necessary to make a decision about the tax proposal. Whether we ought to try to eliminate a particular soft constraint, or try to figure out what to do given the assumption that that soft constraint will remain in place will depend on the scope of the particular non-ideal project in each case.

Classifying Constraints

There are many ways we could classify the different kinds of hard and soft constraints that a theory might have to include as part of its inputs. Andrew Mason, for example, identifies constraints deriving from “historical circumstance, institutional design and human nature”, and describes as a constraint Rawls’ requirement that an adequate theory of justice must be “realistic” (Mason, 2004). Mason also considers constraints based on stability, demandingness, publicity, the potential fallibility of our predictions about what is and is not possible, as well as “feasibility constraints” more generally. Similarly, the kinds of facts that Colin Farrelly thinks a theory should be sensitive to, might be described in terms of constraints upon a theory. Farrelly’s constraints include things like non-compliance, unfavourable social, historical and economic conditions, fallibility, disagreement and human nature (Farrelly, 2007, p.847).

While these are undoubtedly examples of constraints that may or may not apply to an adequate theory of justice, there is no attempt to categorize them, even though some of these constraints are clearly quite different from the others. Requiring that a theory of justice issue recommendations that satisfy a publicity criterion, for example, entails appealing to a substantive normative theory of justification. Requiring that a theory of justice issue recommendations that are not overly demanding requires appealing to an account of morality
more generally. These kind of normative requirements differ from the kinds of practical requirements based on accounts of unfavourable social, historical and economic conditions, or accounts of human nature, or of our epistemic limitations in relevant domains of knowledge, and so on.

I suggest categorising constraints as follows: Constraints may be (1) logical, (2) physical, (3) epistemic, (4) motivational, or (5) moral. With the exception of logical constraints, and depending on one’s metaethical commitments when it comes to the status of moral constraints, each of these types of constraint may come in both hard and soft varieties. The advantage in thinking about constraints in this way is that this allows us to easily categorise the different kinds of project that must be completed in order to produce a theory capable of issuing action-guiding recommendations that are achievable and desirable. Different kinds of constraints will require different skills and resources to identify, assess, and incorporate into the broader theory. This will hopefully become apparent, following some examples of each type of constraint.44

Logical constraints are just facts about the laws of logic that any theory must incorporate if it aspires to offer us coherent claims or conclusions. I take it that such constraints are present at any level of philosophical analysis (including A-Type approaches). A theory that fully identifies and incorporates logical constraints is simply a theory that is not vulnerable to objections based on the logical structure of its arguments. So, for example, if it turned out that some particular conclusion from some particular theory depends on a claim which entails both A and ¬A, if we take the perspective of the CBV, we would say that that theory had failed to identify and incorporate a relevant logical constraint. Identifying these kinds of constraints will normally be fairly straightforward, in that we know the sort of thing we need to do to identify whether a view depends on a claim that is logically inconsistent. All logical constraints are obviously hard constraints.

Physical constraints refer both to the laws of physics, as we currently understand them, and to a broader class of physical facts more generally which concern the resources we have available and what it is possible to do with them. Hard physical constraints track facts about the laws of physics that explain why an agent cannot physically accomplish some particular

44 What is important here is not so much that we use a particular system of categorising constraints, such as the one I describe in this chapter, but rather that whichever categories we use are used clearly and consistently in a way that helps us to understand the relationship between the constraints in question and the ways in which they affect the theory of justice we are working with.
task. While our understanding of physical laws may change, the labelling of a physical fact as a hard constraint should be considered subject to a reasonableness criterion similar to that which I suggested should apply to the labelling of constraints as hard constraints in general: it’s not generally considered a worthwhile objection to a theory to point out that it relies upon the assumption that human beings cannot travel backwards in time, for example, or that five units of an indivisible resource can be distributed equally among six people. Common sense should be our guide in these sorts of disputes when it comes to deciding whether a physical fact should be considered a hard or soft constraint. Soft physical constraints might apply in cases where, for example, a person lacks sufficient resources to accomplish some particular task, but those resources could be made available to that person at some point in the future. Identifying, assessing and incorporating physical constraints into our theory will presumably require a lot of empirical work in addition to the normative work that goes into the theory. To be properly sensitive to physical constraints will often require that the theory take into account relevant facts about the physical resources which are available, the means of acquiring more resources which are available, the costs (both normative and non-normative) associated with acquiring and spending resources, and so on. A theory that fails to incorporate such facts is unlikely to produce the kinds of recommendations that we want a P-Type theory to produce.

Epistemic constraints track the knowledge available to the agents for whom the theory is intended to apply (this may include the theorist herself). Hard epistemic constraints are those facts that it is impossible for the relevant agents to access, while soft epistemic constraints refer to information that is presently unknown to the agents but which could potentially be acquired. These facts can include normative as well as descriptive facts.

For example, suppose that we believe that preference utilitarianism is true, and that what justice requires is simply that the state distributes resources in a way that maximises total preference satisfaction. Upon developing our utilitarian theory of justice, it turns out (predictably) that this is no simple task, since it is extremely expensive to investigate and monitor each individual’s preferences, and to calculate all of the effects of various potential patterns of distribution on people’s preferences. In this case, we are prevented from implementing our theory because the recommendations it issues (i.e. maximise total utility) depend on us having epistemic access to facts which are beyond our reach. Some of these facts will be inaccessible no matter what we do, for example because we simply don’t have enough resources to access them. An example of such a fact might be something like “What
will the precise consequences of this distribution policy upon citizens living in one thousand years’ time?” These are hard epistemic constraints. Some of these facts will be accessible to us here and now, provided that we take steps to access them, which in this case might involve things like surveying the populace, holding referenda or elections, or other means designed to discover what preferences people have. For examples of moral facts that may be epistemically inaccessible, suppose that we are not certain whether a particular species of animal is capable of holding preferences, and thus whether they are a being who is worthy of consideration as part of our theory of justice. If further research into the species’ cognitive capacities would be likely to tell us whether it is capable of holding preferences, then this problem (not knowing whether there are moral constraints with regard to how we should treat these animals) consists of a soft epistemic constraint. On the other hand, suppose that some other theorist is convinced that, while there is a fact of the matter, there is no way to know whether preference utilitarianism is itself correct. For that theorist, they will view the inaccessibility of knowledge of the correct moral theory as a hard epistemic constraint that applies to a moral fact.

Because epistemic constraints can apply to theorists as well as theories, a good theory of justice will need to incorporate the potential for fallibility when identifying and assessing all relevant constraints. I shall return to this issue in much more depth in Chapter 4.

Motivational or psychological constraints are those facts that determine the attitudes and dispositions it is possible for the subjects of a theory to hold, or what it is possible for those subjects to be motivated to do. These kinds of constraints can be especially difficult to identify and assess, given that it is unclear whether there are any hard motivational constraints at all. Some philosophers, notably David Estlund (2011), have even argued that if there are things we cannot be motivated to do, that these should nevertheless not count as a constraint upon the demands of justice.\footnote{I offer arguments against this view in ‘Justice for Jerks: Human Nature, Selfishness and Non-Compliance” [Unpublished Manuscript].} Regardless of whether we agree with that claim, even Estlund would agree that when trying to produce action-guiding recommendations, there are facts about people’s motivations which we must take into account, even if the result is only to produce recommendations that people work to change their current motivations.

The most obvious example of a motivational constraint in the literature comes from the Rawlsian view of ideal theory as ‘full compliance’ theory and non-ideal theory as ‘partial
compliance’ theory. In order for the condition of full compliance to be satisfied, the subjects of the theory must know what justice requires (a condition we can express as a lack of particular epistemic constraints) and must be sufficiently motivated to comply with what justice requires, a condition we can understand as a lack of particular motivational constraints.

Controversy typically arises when a theory of justice seems to produce recommendations that depend on motivations that people may not be able to hold or motivations that we think people ought not be required to hold, or which would be supererogatory to hold. For example, some people believe that cosmopolitan theories of global justice that suggest that we have strong moral obligations to distant strangers are flawed because human beings are simply unable to care about distant strangers as strongly as they care about their co-citizens. For another example, some people believe that a moral theory which requires us to be ‘saints’, perhaps by constantly sacrificing our own personal projects in order to maximise utility is defective because it requires us to act in ways which are either impossible or supererogatory.

Finally, moral constraints are those facts that rule out some particular political proposal on the basis that it would require us to violate a sufficiently weighty moral duty. It is important to be as clear as possible when describe what moral constraints are, and the role that they play in producing political proposals, according to the CBV.

The basic idea is as follows. Before trying to figure out what a particular value like justice requires of us here and now, a theorist will already have some prior commitments in terms of whatever broader moral theory it is to which they subscribe. These initial assumptions about morality are constraints on what is morally permissible, rather than what is physically possible and will be necessary in order to test the intuitive plausibility of the theory’s ultimate recommendations. So, for example, we might take it as a moral constraint that slavery is never justified, and use this assumption to show that a particular political proposal is flawed because it requires the institution of slavery. In other words, a complete and accurate account of non-moral hard and soft constraints will allow us to identify the set of political proposals that it would be possible to implement. A complete and accurate account of moral constraints will allow us to identify the subset of possible political proposals that are also morally desirable, at least with respect to the value of justice.
It is important to note that this idea of moral constraints is neutral with regard to questions of how we can justify our ethical beliefs in general. For example, one could adopt what is called a ‘foundationalist’ view about ethical justification, such that there is some subset of ethical beliefs which are self-evident, or immediately justified, in which case moral constraints would represent this subset of beliefs about morality in a fairly straightforward way.

According to other views about justification, like those of Rawls, while we may have initial ‘fixed points’ when developing our beliefs about morality, virtually everything is ultimately ‘up for grabs’ in the sense of being open to revision in order to maintain a coherent set of views as part of a process of reflective equilibrium (Rawls, 1999, p.18). Adopting this view might lead to several ‘drafts’ of a theory of justice, as we compare our ultimate results with our initial assumptions about morality. Yet in each draft, we must begin with some set of initial assumptions in order to work out what the implications of those assumptions would be. Thus moral constraints will have a similar role to play regardless of whether we opt for a ‘single draft’ or ‘multiple draft’ approach.

Finally, it is important to highlight the fact that a P-Type theory of justice that takes all relevant moral and non-moral constraints into account will only tell us what we ought to do with respect to the value of justice. If we think that the value of justice has absolute priority over all other normatively significant values, then a P-Type theory of justice will be sufficient to produce action-guiding recommendations, since it will be sufficient to tell us how we can make the world a better place with regard to the value of justice. Whether we think justice has this kind of priority will depend upon the conclusions we reached from our theory of ideals, where the aim is to describe what justice is, why it is valuable, and how its value compares to other values. (Of course, depending upon how one describes justice at the level of the theory of ideals, it may be that the right way to understand justice itself is not as a distinct value but as a value that emerges from the right balancing and trade-offs between other distinct values.)

The picture becomes more complicated, however, if we believe that a value like justice does not have absolute priority over other values. In that case, a P-Type theory of justice (assuming that justice is a distinct value) will not be sufficient to produce action-guiding recommendations for us here and now, because we will not yet have tested whether making the world a more justice place via some particular proposal would require an unacceptable trade-off with regard to some other important value. In these cases, we may need to formulate
P-Type theories of these other values, and then to compare the results of these theories with the results of our P-Type theory of justice. Articulating a final action-guiding recommendation under these circumstances will require balancing the recommendations of our theory of justice with our theories of other values, like freedom and equality, in a way which is consistent with our conclusions (via our theory of ideals) about how these values ought to be traded off against one another. One way to do this would be to construct additional moral constraints for our theory of justice which are the result of our (A and P-Type) theorising about other values. So, for example, suppose that our theory of ideals tells us that an increase in the justness of a world is not always worth the corresponding loss of freedom, and that our P-Type theory of freedom allows us to identify all of the ways in which we could make the world a better place with regard to that value. Given the range of possible options identified by a P-Type theory of justice, we might then want to further constrain this set not just by the initial hard moral constraints we began with, but by the additional moral constraints we have discovered via theorising about the relationship between the values of justice and freedom. So, to give a crude example, if we suppose that every one unit of justice is as valuable as five units of freedom, we could use this assumption to rule out any proposals identified by our P-Type theory of justice that would result in a loss of greater than five units of freedom for every one unit of justice gained. Obviously this is an extremely simplistic example, but it should suffice in order to show the ways in which the conclusions we arrive at via theorising about values other than justice could form the basis of moral constraints in a subsequent theory of justice (or indeed, in a subsequent ‘redraft’ of a theory of justice, depending on which value we investigate first).

Each of these categories – logical, physical, epistemic, motivational, and moral – represents a different kind of obstacle that can stand in the way of a political proposal intended to make the world a more just place. In order for a proposal to be achievable, it needs to be logically possible, we need to have (or be able to access) the resources necessary to implement the plan, we need to know what to do with those resources, and we need to be sufficiently motivated to act on the plan with the resources available. In order for the proposal to be desirable, it must be the case that the available plan would result in a more just world, and that the plan would not require us to violate any sufficiently weighty moral duties.

I suggested previously that a ‘pure’ ideal theory would be one that assumes all and only relevant hard constraints as part of its inputs, while a ‘pure’ non-ideal theory would be one that assumes all and only relevant hard and soft constraints as part of its inputs. Between
these two extremes, we can imagine a spectrum of theories that take some soft constraints into account but not others. What would be the point of such an approach? Suppose, for example that we compare the status quo and an ideal state of affairs, with regard to some particular issue of injustice, and discover that there are three obstacles which stand in the way of achieving the ideal which we judge to be a soft epistemic constraint, a soft motivational constraint, and a soft physical constraint. Suppose that the soft physical constraint requires complex technical knowledge to tackle (say, because it requires developing a more efficient method of extracting some important resources), and that we believe this task is best suited to those with the relevant expertise. In a case like this, we might want to proceed as though the physical constraint has been removed, in order to figure out what we then ought to do with regard to the remaining epistemic and motivational constraints. Imagining scenarios where some soft constraints have been removed but not others looks like it will be an important part of developing recommendations about transitioning from the status quo to a more ideal state of affairs. I discuss this idea in much more depth in Chapter 5.

This is not to suggest that every possible approach along this spectrum will be useful or reasonable – it is important for theorists to be clear with regard to which inputs they are including and why. It is also important that when choosing an approach which lies between the two ends of the spectrum, that theorists are careful to be aware of the relationships between different kinds of constraints, given that tackling one soft constraint can often have consequences for our ability to tackle other relevant constraints as part of the same project.

In the next section, I will compare the CBV to the alternatives I outlined in Chapter 1, but before proceeding further, a few final clarificatory points are necessary.

First, whether or not something is to be considered a hard or soft constraint will depend upon the particular context to which the theory is being applied. So, for example, an epistemic constraint may qualify as a hard constraint if there is no possible way that the subjects of the theory could have access to the information (not if there is no possible way anybody could have access to the information). What is relevant is whether or not a constraint is unalterable within the confines of the theory in which it applies. This is one reason why it is vital for theorists to be clear about who their theory is supposed to apply to, and what exactly it is supposed to produce.

Second, it is possible, given the above, that a soft constraint may become a hard constraint over time. An illness may be reversible until it reaches a particular stage, for example, or a
population that might have been controlled may become so large that there are no longer enough resources to go around. This move is unidirectional however; while a soft constraint may become a hard constraint over time, a hard constraint will never (or can never reasonably be expected to) become a soft one (or at least, I can think of no possible example where this might be true).

Third, it will sometimes be impossible to know whether a particular constraint is a hard constraint or a soft constraint until we try to remove or minimise it. A disease may turn out to be beyond our capacities to cure, for example, but we might not discover this until we have invested significant amounts of time or research into attempting to cure it (assuming we ever get to the point where we could prove that it is beyond our ability to ever cure).

Finally, when it comes to the real world, constraints may well involve combinations of hard and soft constraints, and involve combinations of physical, normative, epistemic and motivational constraints. Suppose for example, that a particular disease cannot be cured, but that its effects can be lessened such that the lives of sufferers are extended by a considerable period. In this case there is an underlying hard constraint that prevents a cure, but on top of this are softer constraints that can be remedied through the appropriate medicine. But then suppose that administering this medicine would be so expensive as to require the withdrawal of funds that would otherwise have been spent on saving the lives of a great many more patients in which case normative constraints may also apply.

3.3 – Comparing the Constraint-Based View to the alternatives

Having described the CBV in the previous section, I now turn to the task of arguing for the advantages of the CBV over alternative ways of conceptualising the structures of, and relationship between, ideal and non-ideal theories. In Chapter 1, I outlined some of the more popular conceptions of ideal and non-ideal theory. In this section, I will compare the CBV to these alternatives and argue that we have good reasons in each case to prefer the CBV as a way of conceptualising ideal and non-ideal theory.
(i) The CBV and the Rawlsian View

Recall that on this view, a theory is said to be ideal when it assumes full compliance, i.e. if it ascribes certain capacities, knowledge and motivations to those for whom the theory is intended to apply.

Rawls tells us that he is primarily concerned with examining the principles of justice which would regulate a “well-ordered society” which he defines as a society in which “(1) everyone accepts and knows that the others accept the same principles of justice, and (2) the basic social institutions generally satisfy and are generally known to satisfy these principles” (Rawls, 1999, pp.397-398).

To know and accept a principle of justice will require in the first place that people are free of epistemic constraints that would otherwise prevent a theory from being comprehended. This, for example, is why theories of justice are not generally addressed to very young children, who could not reasonably be expected to understand complex theories in the first place, let alone endorse them. Since mental immaturity of this sort is a hard constraint, an ideal theory of justice cannot simply assume it away. Thus Rawls acknowledges that:

> even in a well-ordered society under favourable circumstances...[liberty may be justifiably restricted in response to]...the more-or-less permanent conditions of political life. [Other restrictions] are adjustments to the natural features of the human situation, as with the lesser liberty of children. In these cases the problem is to discover the just way to meet certain given limitations. (Rawls, 1999, p.215)

These limitations – the “more-or-less permanent conditions of political life” and “the natural features of the human situation” are plausibly understood as the set of those facts that constitute all relevant hard constraints, according to the CBV. Consider also the following passage from A Theory of Justice:

> But once the ideal conception is chosen, [parties in the Original Position] will want to insure themselves against the possibility that their powers are undeveloped and they cannot rationally advance their interests, as in the case of children; or that through some misfortune or accident they are unable to make decisions for their good, as in the case of those seriously injured or mentally disturbed. (Rawls, 1999, pp. 281-219)

Here we can see quite explicitly the role that some hard constraints play in Rawls’ project. Children will always be constrained, both physically and intellectually, and there is nothing
we could ever reasonably hope to do to change that fact. Similarly, we can see here some other assumptions made by Rawls about the constraints that will apply to us even in an ideal world: we may be subjected to accidents or have the misfortune of being born with or coming to develop physical and intellectual disabilities. As can be seen above, the assumption that these constraints will apply is a direct motivating influence to the parties in the original position.

Returning to the condition of full compliance and of the possibility of a well-ordered society, it is not merely sufficient that the subjects of the theory can and do understand the principles of justice and that they can know whether or not their institutions are working in accordance with the principles. They must also be free of motivational constraints that might prevent them from acting upon the theory even though they understand it and acknowledge it to be correct.

Additionally, in order for the social institutions to satisfy the principles of justice, resources will be required to establish and maintain those institutions. This explains Rawls’ assumption of “favourable conditions”, and can be thought of as involving the absence of various physical constraints (or, conversely, in terms of the presence of various physical capacities). The assumption is that we will have the resources required to establish and maintain whatever institutions turn out to be demanded by the principles of justice.

In their discussion of the Rawlsian view, Hamlin and Stemplowska note that compliance comes in different types, and can be measured on a spectrum, depending on how prevalent compliance is. They describe the assumption of full compliance as entailing that “almost everyone does almost everything that the normative content of that theory demands of them”. Understood in this way, we can imagine a range of different kinds of compliance, depending on how many people comply, and the extent to which those who comply do so (Hamlin & Stemplowska, 2012, pp.49-50).

These further complications can be accommodated relatively straightforwardly by the CBV: if we want to measure compliance in all of its dimensions, then we must measure the strength of multiple kinds of constraints. This renders the constraint-based view compatible both with the idea that the ideal/non-ideal distinction is best-conceived as multidimensional (since it tracks multiple kinds of constraints) and with the idea that it is best conceived as a continuum (since constraints can be analysed according to their strength and according to their pervasiveness).
So, for example, if we want to know why a theory is not widespread, the constraint-based view would suggest comparing the constraints that apply to those who do accept the theory to the constraints that apply to those who do not. We might discover, for example, that a theory is only held by a small number of well-educated people, owning to the absence of certain epistemic and physical (resource-based) constraints which tend to apply to most other people.

Similarly, when examining how closely individuals comply with a theory, we should be able to explain differences in levels of individual commitment to a theory by talking about differences in the constraints that apply. Perhaps one group of people will turn out to be less committed to some principle of environmental justice because they lack the physical resources to build recycling plants, for example.

To conclude, the CBV is capable of accommodating the Rawlsian view, and may even offer greater clarity when dealing with issues of compliance, given that the CBV breaks compliance down into its different parts and dimensions by describing compliance in terms of the various physical, epistemic and motivational constraints that apply. The flaw with the Rawlsian view, however, is that it is too narrow: there is more to ideal and non-ideal theory than just full and partial compliance.

(ii) The Constraint-Based View and Idealisation and Abstraction

Idealisations and abstractions are different kinds of assumptions we make when constructing a theory. The purpose of both assumptions is to serve a ‘simplifying’ function, making it easier for us to draw conclusions when dealing with complicated questions. Abstractions simplify by ‘bracketing off’ facts about the world that may then be reintroduced at a later point without altering the conclusions we have drawn. In contrast, idealisations are assumptions that are actually false. According to Charles Mills’ criticism of ideal theory, ideal theory is distinct in that it involves idealisations rather than mere abstractions, idealisations are likely to change the outputs of the theory, and ideal theory is flawed for that reason.

One way to respond to this objection is to question the usefulness of the distinction between idealisations and abstractions. Another way to respond to the objection is to accept the distinction between idealisations and abstractions, but to deny that ideal theorists use (or need to use) idealisations in order to engage in ideal theory.

Recall how the CBV characterises ‘pure’ ideal theory for example. According to the CBV, we engage in this approach when we include as part of the inputs to our theory all and only relevant hard constraints. These, by definition, are facts about the world which we have good reason to think cannot ever be changed. So, they are not abstractions, since they do not represent simplified facts, nor are they idealisations, since they do not represent false assumptions. According to the CBV, no other factual assumptions are included at this end of the spectrum of ideal to non-ideal theories.

So, where are the idealisations to be found in such an approach? To identify an idealisation, we must be able to point to an assumption made by the theory (either explicitly or implicitly) and then show that this assumption is false in the real world.

Presumably, the relevant assumptions are to be found in the fact that a purely ideal approach assumes that soft constraints do not exist. This will almost always be a false assumption, and so we might think that these are the harmful idealisations that will distort the conclusions of our theory.

Yet this way of putting the objection only seems to work if we assume (and this is certainly a false assumption) that ideal theorists will attempt to apply the conclusions of their ideal theories directly to the non-ideal world. The pure ideal theory in question will be sufficient to generate recommendations for a world that conforms to its assumptions – in other words, a hypothetical world which does not (and which is not assumed to) exist. Should we say that such an approach involves idealisations, even though such an approach is not intended to offer a model of the real world?

Even if we do concede that these kinds of assumptions should still count as idealisations, it seems clear that these sorts of assumptions will not in fact make us more likely to arrive at the wrong recommendations when it comes to how we ought to act in the real world. This is because, according to the CBV, we cannot issue action-guiding recommendations for the real world until we have taken into account all of the relevant soft constraints as well as all of the relevant hard constraints. As we move along the spectrum towards the non-ideal, we incorporate more relevant soft constraints, and the world we are modelling changes with each step. It is not possible, therefore, for us to issue action-guiding recommendations about the real world, which are based on false assumptions, except if we have made an empirical error somewhere along the way.
While we should therefore reject Mills’ criticism of ideal theory in general, it is important to be aware of the potential for a theoretical process to become distorted if the wrong assumptions are made.

If we think in terms of constraints, what sort of errors might lead us to issue recommendations based upon assumptions that are actually false? There are two kinds of false assumptions that are suggested by the constraint-based view, which might lead to this problem. The first would be false assumptions about constraints in cases where we falsely assume that a relevant constraint upon our behaviour is not in fact relevant to the theory. A theory that ignored moral constraints to do with political legitimacy, for example, might end up recommending a benevolent dictatorship on the basis that the leader in question would pursue policies that would be in the best interests of his people. This problem of a failure to identify a constraint can apply either in cases where the constraint is not recognised as being relevant to the issue under consideration (as in the above example) or in cases where we falsely assume that a particular fact constitutes a relevant constraint.

The second kinds of errors arise from what we can call the problem of identification. This can occur even when we recognise the relevance of a particular constraint but erroneously determine it to be a hard constraint rather than a soft constraint or vice-versa. It can also happen if we fail to properly gauge the strength of the constraint in question. For example, if our theory says that everyone ought to become a vegetarian, but that the chief constraint against this is the price of fresh fruit and vegetables, we might be motivated to take actions designed to make the monetary costs of vegetarianism less burdensome. If, however, it turns out that most people are put off becoming vegetarian not by the cost of food but by the erroneous belief that a diet that does not include meat is not healthy, then we would likely change our strategy to deal primarily with the soft epistemic constraints rather than the soft physical constraints which prevent people from becoming vegetarian.

Assumptions that treat hard constraints like soft constraints are likely to present obvious problems for a theory. Resources may be wasted trying to remove obstacles that cannot possibly be removed, demands may be placed upon people that are overly burdensome, and we may well be left worse off as a result. On the other hand, assumptions that treat soft constraints like hard constraints may, at the very least, result in outcomes that are far less desirable than what we could achieve by taking action to minimise or eliminate the soft constraints in question.
If a theory achieves completely accurate identification of all of the relevant constraints that apply, then we should not be worried about the resulting conclusions being based upon idealised assumptions that are false. The relevant distinction here seems to be a distinction between a theory that has properly identified all of the relevant constraints (and identified the relative strength of these constraints), and a theory that has failed to recognise relevant constraints, mislabelled constraints as being hard when they are soft (or vice versa), and/or misjudged the relative strengths of the constraints it identifies.

To summarise this subsection, we should reject the characterisation of the difference between ideal and non-ideal theory as the difference between a theory involving idealisations and one involving abstractions. If we adopt the CBV, we can see that either ideal theory need not (and should not) involve idealisations, or that the kinds of idealisations it does require will not be problematic, contrary to Mills’ claims.

(ii) The Constraint-Based View and Fact-Sensitivity

The third view about the relationship between ideal and non-ideal theory that I described in Chapter 1, as set out by Colin Farrelly, suggests that ‘ideal theory’ is the same as ‘insufficiently fact-sensitive theory’. In Chapter 2, I argued that at least some of what Farrelly considers to be ideal theory, i.e. A-Type approaches, which do not aim to (ultimately) produce action-guiding recommendations, ought to be distinguished from those that do, i.e. P-Type approaches.

This leaves us with a set of P-Type approaches that would be divided, on Farrelly’s view, between sufficiently and insufficiently fact-sensitive approaches. This distinction forms the basis of Farrelly’s criticism of those approaches (both A-Type and P-Type, since Farrelly does not recognise that distinction) which he holds to be insufficiently fact-sensitive.

If the constraint-based view is the right way to understand the relationship between ideal and non-ideal theory, we can show that Farrelly’s criticism of P-Type theory is either misguided or mistaken. It is misguided if the objection is supposed to be that particular ideal theories tend not to recognise certain important hard constraints. If that is the objection, then the objection is not to ideal theory in general, conceived as the constraint-based view conceives it, but rather to particular instances of ideal theory that are straightforwardly defective on their own terms because they fail to take into account what they should be taking
into account. Or, if the objection is supposed to be that ideal theory fails to give us action-guiding recommendations, then it is misguided, since no ideal theorist thinks that ideal theory is sufficient for that task. It is only once we start to add soft constraints into the mix that we get closer to issuing action-guiding recommendations via the identification of political proposals that are achievable and desirable. Farrelly might reply that there is something about beginning with the ideal that means that subsequent considerations are somehow distorted or infected with faulty initial assumptions, but the constraint-based view shows that this fear is unfounded. Ideal theory only includes those assumptions that must be present at any level of theorising (with the exception of the theory of ideals). It is not possible, therefore, for ideal theory to “distort” non-ideal theory with faulty assumptions, since non-ideal theory will have to include these assumptions in any case.

We ought then to salvage from Farrelly’s position a distinction between fact sensitivity and fact insensitivity that does not map on to a distinction between ideal and non-ideal theory, but which rather maps on to a distinction between “good” theory and “bad” theory. A “good” theory will be one that takes all of the appropriate facts into account, and a “bad” theory will be one which fails to, such that its failure to do so impacts upon the outcome of the theory. The constraint-based view is capable of distinguishing between the kinds of facts appropriate for different kinds of theories—hard constraints only for one end of the ideal/non-ideal spectrum, hard and soft for the other end of the spectrum, and no constraints for theories that fall into the axiological category. The above is still compatible with the multidimensional continuum view since there will be different kinds of facts/constraints to take into account and a theory can be measured according to how many facts it takes into account (ranging from a theory which fails to take into account most of the facts it ought to, to a theory which takes into account almost all the facts it needs to, as far as we can discern).

(iv) The Constraint-Based View and the Transcendental/Comparative distinction

Amartya Sen’s distinction between transcendental and comparative approaches to justice may seem to fit the distinction between ideal and non-ideal approaches, though it must be noted that Sen himself never explicitly equates the two. Simply put, we engage in the transcendental approach when we are concerned with investigating the features of a perfectly just society, and we engage in the comparative approach when we are concerned with investigating the relative justness of any two (or more) possible societies.
Since Sen identifies the transcendental approach with theorists like Rawls, we can interpret a “perfectly” just society as resembling a Rawlsian realistic utopia, which on the constraint-based view is a world where only hard constraints apply to the subjects of the theory. It might seem straightforward to then identify Sen’s transcendental approach with ideal theory (which, on the constraint-based view, is precisely any theory which is concerned with a society in which only hard constraints apply). But there is another feature of Sen’s distinction that makes this identification problematic. This concerns Sen’s assumption that a perfectly just society is necessarily unachievable such that comparative judgments are assumed to be judgments between different viable states of affairs, and only less-than-perfect states of affairs are assumed to be viable.

It might be possible to engage in ‘pure’ ideal theory (as hard-constraints-only theory) without concern for viability, though this would be a strange project to engage in intentionally since if viability is not our concern there is surely little reason to pay attention to any kind of constraint in the first place and we would be better off engaging in an A-type approach. But as we have seen, the Rawlsian brand of ideal theory is explicitly concerned with questions of viability in the long term since it is motivated by the problems that present themselves in non-ideal theory and sets itself up as step one of a two-step process towards fully addressing these problems. Hard constraints, which Rawls aims to take into account, are inextricable from notions of viability, as are all constraints more generally.

Consider also cases that involve not a just society as a whole, but some specific issue of injustice within a society. It’s reasonable to suppose that there are some injustices we have been successful at wiping out in certain contexts, say slavery in the United States, for example. If we were to construct an “ideal theory” picture of what a just society would look like with regard to slavery, we would find that the transcendental account and the status quo are essentially one and the same. When we consider specific issues within a society, there is no reason to suppose a priori that the answer we get from the question “what would a perfectly just society do about this?” might not be identical to “what can we do about this here and now?” To put this in terms of the constraint-based view, sometimes it will simply be the case that no constraints, hard or soft, prevent us from meeting the demands of justice in some particular sphere. Furthermore, the extent to which ideals tend to be achievable will presumably depend on the precise content of the theory of justice we endorse. Implementing a highly complex set of liberal distributive requirements in a highly complex society would
presumably be a much more difficult task than, say, implementing a minimal libertarian state in a small isolated community.

One might insist that Sen’s categories only apply to considerations of societies as a whole, and that it would be highly implausible to think that we are anywhere near capable of bringing about a perfectly just society given the world we find ourselves in at present. The first point to note in response is that implausibility is not impossibility and that (at least, theoretically speaking) there is no necessary connection between a state of affairs being the best imaginable and it being unachievable. A stronger reply to this, however, is to reject the notion that we can only engage in ideal theory approaches when we are considering societies as a whole. Certainly, on the constraint-based view there is no reason why we cannot take this approach in dealing with specific sub-domains of justice. While eradicating slavery, or the death penalty, or any other reform may have consequences that impact upon society more generally, it would be very strange to think that we would first need to transform society as a whole rather than chip away at particular issues of injustice which we can eradicate or minimise here and now, provided that we have sufficient reason to think that the normative costs for such actions are acceptable.

Indeed, this is precisely Sen’s point- we do not need a full and complete account of a perfectly just society in order to improve the status quo with respect to some particular cases of injustice. Sen, however, appears to assume that because we can make these kinds of improvements now, ideal theory in general is neither necessary nor sufficient to affect change in the real world. This assumption may be warranted if ideal/transcendental approaches are assumed to be unviable; if we can do something about injustice now, and transcendental theories give us nonviable recommendations (or do not lead to viable recommendations at any point), then it might appear that they play no part in our subsequent actions. But it is a mistake to assume anything about the viability of a theory’s recommendations merely because it only takes hard constraints into account, and especially if that process is intended (as P-type approaches are) to be the first part of a multi-stage process.

In contrast to the transcendental approach, comparative judgments are assumed to be made between different viable states of affairs. Assuming that we have a sufficient understanding of viability in the first place that we can correctly identify our options as being actually achievable, we require a further understanding of desirability in order to allow us to choose between options. But where exactly is this understanding to come from?
Consider a case where we have two possible societies, identical in every respect except for the fact that Society A tortures political dissidents and Society B does not. In order to determine that we should bring about B rather than A we need, for example, a theory that tells us that torturing dissidents is wrong. In this case, we could imagine a perfect society (C), observe that this society would not torture its dissidents, and use this as the basis for choosing to bring about B rather than A. In this case, using ideal theory (by imagining the kinds of constraints that would be present in a perfectly just world, and noting that there would be a moral constraint against torturing political dissidents) seems a perfectly sufficient way for us to make the comparative judgment that B would be a more desirable society than A.

Is ideal theory necessary in order to make comparative judgments? It may not be, in the sense that we could begin our theorising by considering a range of constraints without distinguishing between hard and soft constraints. In this case, however, a failure to distinguish between hard and soft constraints will make it much less likely that we are able to (1) identify the range of political proposals that are actually available to us (since we may mistakenly treat hard constraints as soft constraints) or (2) choose the best option from those that are available to us (since we may fail to appreciate the relative strengths of the relevant constraints). It is possible, of course, that we might simply get lucky, and choose the best available option without having identified and assessed the relevant constraints, but if ideal theory is not necessary for these kinds of judgments it seems to be highly prudent to begin with it, at the very least.

It seems clear that the transcendental/comparative distinction just won’t do as a viable conception of the distinction between ideal and non-ideal approaches. This would force us to conflate “ideal” with “necessarily unviable” and to reject the idea that ideal approaches can or should play a part in non-ideal considerations. This puts the view squarely at odds with the constraint-based view where viability is not necessarily connected to the kind of theory one is producing and, more importantly, where all non-ideal theory is assumed to have at least some kind of minimal foundation in ideal theory. Hard constraints can be considered on their own, especially as the first step in a process that moves along a spectrum towards less ideal considerations, but all non-ideal theorists will have to consider both hard constraints and soft. This presents us with the plausible assumption that non-ideal theorists will tend to produce recommendations that will probably be more constrained and limited in scope (since they are taking more constraints into account) than those which might be produced by those ideal theorists considering hard constraints alone. It’s certainly not compatible with the
transcendental/comparative distinction which would suggest the bizarre view that (if transcendental/ideal theory is not necessary for comparative/non-ideal theory) non-ideal theorists should produce theories that are somehow constrained by facts about the world that we can change, but unconstrained by facts about the world that we cannot.

(v) The Constraint-Based View and the Theory of Ideals

On Hamlin and Stemplowska’s view, the theory of ideals refers to a type of project that is distinct from the multidimensional continuum of ideal and non-ideal theories of justice. The distinction between the two can be understood in terms of the purpose of the project – when we engage in the theory of ideals, we are aiming to produce recommendations about what we should think about certain concepts and/or what we should think about the relationship between different concepts or values when it comes to things like questions of prioritisation or trade-offs.

Expressed in terms of constraints (or rather, a lack of constraints), the theory of ideals seems to me to be more or less equivalent to A-Type theorising as described in Chapter 2. Recall that A-Type theorising is characterised by two features in particular – its aims and its inputs. Specifically, A-Type theorising aims to produce conceptual clarity about the nature of concepts or values that are relevant to the pursuit of justice rather than recommendations about what we ought to do to make the real world a more just place. A-Type theory also requires the theorist to be constrained only by basic rules of logic that constrain any philosophical activity. Introducing further hard and soft constraints moves us away from A-Type theory and into P-Type theory, or to use Hamlin and Stemplowska’s model, from the theory of ideals towards ideal and non-ideal theory.

This distinction between the theory of ideals and ideal/non-ideal theory is therefore one that is perfectly compatible with the constraint-based view, where it seems to match the distinction between A-Type and P-Type theorising. This distinction is especially important because it offers a strong line of defence for some of those theorists whose approaches have been attacked by critics of ideal theory, as I aimed to show in Chapter 2. Consider the following from Colin Farrelly, for example:

Rather than moving in the direction advocated by Cohen and Mason (i.e. towards a more extreme idealized position) political philosophers should take more seriously non-ideal theory. This will help equip them with a theory of justice that can provide some normative guidance for real, non-ideal societies (Farrelly, 2007, p. 845).
It seems clear from the above that Farrelly conceives of a single spectrum of possible positions with theorists like Mason and Cohen at the ‘extreme’ and theorists such as Rawls as being more ‘moderate’. In the above quotation we can also see evidence of Farrelly’s assumption that the point of this kind of theorising is (or ought to be) to give us “normative guidance for real, non-ideal societies”.

Recall again the critics’ argument outlined in the previous chapter:

- **P1** – We should reject any method of theorizing about justice that fails to, or is not concerned with, providing us with action-guiding principles in the real world

- **P2** - The Rawlsian method (or any variant that relies heavily upon ideal theory) fails to provide us with action-guiding principles in the real world.

- **C** - We should reject the Rawlsian method (or any similar variant that relies heavily upon ideal theory).

If the distinction between the theory of ideals and the ideal/non-ideal spectrum is correct, and if those who engage in the theory of ideals are doing something that is worthwhile, then the first premise above is false. As I argued in Chapter 2, by failing to recognise the distinction, Farrelly misses the point of at least some kinds of theory- those theories that are not aimed at providing “normative guidance for real, non-ideal societies”, but which are instead aimed at investigating the nature of ideals themselves. We might well be suspicious of a theory that is designed to produce action-guiding recommendations and fails to do so, but, as I suggested earlier, there is a difference between a theory that functions as it is intended to, and one that does not.

If the distinction is correct, and if theorists like Mason and Cohen are engaged in the theory of ideals rather than ideal theory, then it’s not sufficient for Farrelly to show merely that they fail to provide us with guidance for the real world. What must be shown is that there are good reasons to reject the sorts of approaches that belong to the theory of ideals in the first place. This will require a different kind of argument to the sort Farrelly aims at Rawls, since both Rawls and Farrelly assume that the purpose of their preferred mode of theorising is to ultimately produce action-guiding recommendations. In Chapter 2, I considered the sorts of arguments that might be made against engaging in the theory of ideals, and attempted to defend such theorists from the charge that their projects are fundamentally flawed.
3.4 Conclusion

This chapter has had two central aims – to describe the Constraint-Based View of ideal and non-ideal theory, and to compare this view to some of the most plausible alternative ways of conceptualising the relationship between ideal and non-ideal theory.

To summarise; the CBV conceives of a (multidimensional) spectrum of approaches, from “pure” ideal theory at one end to “pure” non-ideal theory at the other. Pure ideal theory consists of asking what justice would look like in a world where only hard constraints apply. Hard constraints are those facts about the world that are relevant to the pursuit of justice and which we are confident will never be changed. Pure non-ideal theory consists of asking what justice would look like in a world where we include every relevant hard and soft constraint as part of our inputs, where soft constraints are facts relevant to the pursuit of justice that we have good reason to think we could change.

The CBV includes physical, epistemic, motivational and moral constraints as part of its view, though the precise taxonomy is not essential to the more general view.

Because we can construct approaches that take some soft constraints into account but not others, (for example, by assuming only hard epistemic constraints but including soft motivational constraints) we can produce approaches that are more ideal in some dimensions but less ideal in others.

Comparing this view to the other views described in Chapter 1, I have argued:

(1) That the Rawlsian view is compatible with the CBV, but that the CBV is able to accommodate additional dimensions besides compliance that are worth taking into account.

(2) That the distinction between abstraction and idealisation is not a helpful one to use when considering the differences between more ideal and less ideal approaches. The CBV seems to show either that we can do ideal theory without idealisation, or that the sorts of idealisations we would need to use should not trouble us.
(3) That Farrelly’s distinction between (sufficiently) fact-sensitive and (sufficiently) fact-insensitive approaches is useful as a way of distinguishing between ‘good’ theory and ‘bad’ theory, but not between ideal and non-ideal theory. Assessing Rawlsian approaches, for example, in terms of their aims and the kinds of constraints they take into account, seems to show that they are not insufficiently face-sensitive.

(4) That Sen’s distinction between transcendental and comparative approaches does not plausibly map on to a distinction between ideal and non-ideal.

(5) That Hamlin and Stemplowska’s distinction between the theory of ideals and ideal and non-ideal theory seems to map on to the distinction between A-Type and P-Type theory, and that their view seems to be entirely compatible with the CBV.
4. The Constraint-Based View and Feasibility*

4.1 Introduction

In *A Theory of Justice*, Rawls says that "the reason for beginning with ideal theory is that it provides, I believe, the only basis for the systematic grasp of [the problems of partial compliance theory]" (Rawls, 1999a, p.8). If we follow Rawls, we care about ideal theory primarily because we care about non-ideal theory: we think that an effective way to figure out what we ought to do here and now is to begin by considering what justice would look like in an ideal world.

In terms of the Constraint-Based View, this means beginning with an analysis of the world in which only hard constraints apply and figuring out what justice would look like in such a world, before adding more and more soft constraints to the picture, making the theory less and less ideal. Ultimately, the aim of this process is to produce a theory that takes the world as it is, including all relevant hard and soft constraints, and issues action-guiding recommendations based on that picture of the world.

On this view, ideal theorising marks the first step in a multi-stage process that culminates in the production of what Zofia Stemplowska calls "AD-recommendations", or recommendations that are achievable and desirable for people in the real world (Stemplowska, 2008, pp.323-25). A recommendation of this sort is any plan of action aimed at bringing about a state of affairs that is more desirable than the status quo, to the extent that it represents a more just state of affairs. For the remainder of this chapter, I will refer to these sorts of plans simply as “political proposals”.

It is important to note that there is a distinction between assessing a state of affairs in terms of whether it is achievable and desirable, and assessing a political proposal in terms of whether it is achievable and desirable. For some particular state of affairs (S) to be achievable at some particular point in time (T) it must be the case that there exists at least one political proposal (or set of proposals) which it is possible to implement at T and which could bring

---

47 The order in which soft constraints are ‘added’ may be significant, and we may require a further method of helping us to identify the appropriate order in which to add constraints. I do not have the space to address those complications here, except to suggest that
* A revised version of sections of this chapter has been published as Carey, B. (2014) ‘Towards a ‘Non-Ideal’ Non-Ideal Theory’, Journal of Applied Philosophy, Vol. 32(1).

48 1. By “political proposals” I mean any action(s) intended to bring about a more desirable state of affairs in the political sphere of morality.
about S provided that it is implemented. For S to be desirable, however, S need not be achievable. It merely needs to be the case that S represents a state of affairs which we judge to be evaluatively better than the status quo even if there is no possible way to bring S about. On this view, to say “S is desirable” is to say something like “It would be better if S were the case.” It is not to say “we ought to try to bring S about.” To make a claim of the latter sort is to make a claim about a state’s normative desirability rather than its evaluative desirability. Unless otherwise specified, whenever I shall refer to the desirability of a state of affairs, I am referring to its evaluative desirability. It is only once we move to consider the possible ways of achieving an evaluatively desirable state of affairs that we can assess its normative desirability.

For some particular political proposal (P) to be achievable, it must be the case that it is possible to implement it, or in other words, that it is possible to follow the plan of action it recommends. For P to be desirable it must be the case that implementing it, or trying to implement it, is likely to produce an evaluatively desirable outcome, or satisfy significant moral duties, where these values are understood as being improvements upon the status quo. To say “P is achievable and desirable” and “P is more desirable than any alternatives” is to say “we ought to try to implement P”.

It would be a mistake, therefore, to object to ideal theorising on the grounds that it is not sufficient to produce action-guiding recommendations that are achievable and desirable— if that was what ideal theorists thought, then they would think that we can begin and end with ideal theory, but no defender of ideal theory thinks that. Rather, ideal theorists must defend the claim that their approach is either necessary, or part of a better method than alternative approaches to dealing with what we take to be the pressing and urgent problems presented at the level of non-ideal theory.

If we should think of ideal theory as the first step of a larger process, and if such a process does not tend to produce action-guiding recommendations that are achievable and desirable, we can conclude that ideal theory is not fit for purpose. Since achievability is vital to this process, a defender of ideal theory needs to be able to explain the role that feasibility

---

49 Most obviously, Rawls’ explicit claim that he is “beginning” with ideal theory suggests he views ideal theory as just one step of a larger process, which includes non-ideal theory. See also Stemplowska (2008), Simmons (2010), and Valentini (2012).
considerations play in the process at both the ideal and non-ideal stages. By “feasibility considerations”, I mean the assessments we make about whether it would be possible to implement a particular political proposal, and assessments about a political proposal’s likelihood of success, provided that we try to implement it. When we apply feasibility considerations to states of affairs rather than political proposals, the question becomes whether there are any feasible political proposals that stand a chance of bringing about the state of affairs in question, provided that we try to implement those proposals.

Recent work on the concept of political feasibility by Pablo Gilabert and Holly Lawford-Smith has helped to clarify the sorts of roles feasibility considerations can play in political philosophy in general, and in ideal and non-ideal theory in particular (Gilabert & Lawford-Smith, 2012). In Section 4.2, I draw on this work in order to offer an account of the role that feasibility considerations should play at both the ideal and the non-ideal stage of theorising about justice, according to the Constraint-Based View. On this view, ideal theory should assume all and only those facts that constitute hard constraints upon what it is possible and desirable for human beings to achieve. At this stage, feasibility considerations serve what Gilabert and Lawford-Smith refer to as a "ruling out" function, by eliminating impossible states of affairs from consideration. At the non-ideal level, soft constraints are taken into account, and feasibility considerations take on a scalar rather than a binary function, serving to rank possible political proposals in terms of the probability that they can be achieved, given all of the relevant constraints that apply here and now.

While the account I shall endorse is close to Gilabert and Lawford-Smith's, my view differs from theirs in two ways. Firstly, I include (hard) moral constraints as part of my view, which are facts that allow us to distinguish between desirable and non-desirable states of affairs at the level of ideal theory, and desirable and non-desirable political proposals at the level of non-ideal theory. Second, my view considers feasibility assessments from an objectivist and a subjectivist perspective, in contrast to Gilabert and Lawford-Smith's exclusively objectivist approach. This allows my view to incorporate additional important epistemic constraints that impact upon our capacity to make accurate feasibility assessments.

---

50 To speak of the “level of ideal theory” on the one hand, and the “level of non-ideal theory” on the other may suggest a sharper distinction between the two than I believe exists. I agree with Hamlin & Stemplowska (2012) that it is most accurate to speak of theories as being more or less ideal, depending upon the features of the theory one is considering. In this paper, whenever it seems as though I am drawing a sharp contrast between ideal and non-ideal theory, what I have in mind is a contrast between two ends of a spectrum, with “hard constraints only” theories on one end, and “all relevant hard and soft constraints” theories on the other.
In Section 4.3, I identify two important problems raised by this way of conceptualising the role of feasibility considerations in ideal and non-ideal theory. First, there is the problem of trade-offs: it is not clear how we ought to weigh up the value of a proposal’s feasibility against other values in cases where all other things are not equal. In order to illustrate the persistent and pervasive nature of this problem, I draw on a distinction between ‘destination comparisons’ and ‘route comparisons’, whereby the former involve comparing the feasibility values of a range of mutually exclusive outcomes (or states of affairs), and the latter involve feasibility comparisons among a range of mutually exclusive ways of achieving a particular outcome.

I then examine a related issue that I shall call the problem of fallibility. Here I turn to the issue of subjectivist feasibility assessments and identify certain epistemic constraints that apply to the moral and empirical judgements we make when assessing a political proposal’s feasibility. This problem compounds the difficulties we encounter when considering the problem of trade-offs, because it introduces an additional variable that must be considered when weighing up the value of any particular political proposal.

In Section 4.4, I offer a non-ideal treatment of these problems. I suggest we aim to improve our capacities to make accurate feasibility assessments in both ideal and non-ideal theory such that we can minimise the chances that we will make the wrong decisions when it comes to trading feasibility off against other values in non-ideal theory. To support this ‘narrowing down’ approach, I use an analogy of a chess player to describe a process whereby we come to develop the skills to recognise inferior ‘moves’ (or political proposals) from superior ones.

4.2 – Feasibility and the Constraint-Based View

Describing a world where only hard constraints apply allows us to describe the best possible state of affairs, so one task for the non-ideal theorist will be to describe the circumstances under which we are ever justified in deviating from the pursuit of this ideal. In some cases, we may believe that we should pursue an alternative to an ideal if the chances of successfully achieving that ideal are sufficiently low.
Feasibility in its scalar sense is a property of a political proposal that determines how likely it is that that proposal can be implemented, assuming that we try. Whereas binary feasibility assessments divide possible and impossible proposals, scalar feasibility assessments rank them on a spectrum from “extremely likely to succeed, if we try” to “extremely unlikely to succeed, if we try”. While hard constraints make the difference when it comes to binary feasibility, the presence and nature of soft constraints determine a proposal’s scalar feasibility.

It is not difficult to see the normative significance that attaches to whether a proposal is feasible, though it is much less clear how this value should be weighed up against other normatively significant properties of a proposal. What is clear is that feasibility is not intrinsically valuable; that some political proposal is extremely feasible tells us absolutely nothing on its own about how desirable that proposal might be- a proposal to implement some extremely unjust law might be perfectly feasible, for example, but its being feasible would not imply that it is desirable. However, feasibility does seem to serve a kind of amplifying function in cases where we must choose between proposals that are otherwise similarly desirable. If we are faced with the choice of saving A or B, and if we have a 99% chance of saving A but only a 1% chance of saving B, we ought to try to save A, all else being equal. In this case, it looks like the existing value, or desirability, of one proposal is amplified because it is more feasible.

If the (scalar) feasibility of a political proposal is to be explained in terms of the soft constraints that stand in the way of its implementation then non-ideal theorists must find ways to identify as many relevant soft constraints as possible, and to accurately describe the extent to which they constrain the implementation of the political proposal in question.

Recall that these constraints can be classified in different ways. Gilabert and Lawford-Smith (hereafter G&L-S) suggest categories including logical, physical, biological, economic, institutional and cultural constraints, and possibly psychological and motivational constraints (Gilabert & Lawford-Smith, 2012, p.83). Some of these presumably come only in the ‘hard’ variety, such as logical constraints, while others will have hard and soft varieties. Some kinds of constraint may only be soft - G&L-S believe this to be true of economic, institutional and cultural constraints, for example.

---

51 If a hard constraint prevents a proposal from being implemented then that proposal is straightforwardly infeasible in the binary sense.
Perhaps the key distinction between hard and soft constraints made by G&L-S is that the latter are malleable in a way in which the former are not. Drawing on Mark Jensen's concept of “indirect diachronic ability” (Jensen, 2009) they point out that we can be constrained by facts which it is possible for us to change, provided that we take certain steps. If this is correct, then we should perhaps say that any kind of constraint can potentially be a hard constraint. Malleability looks like it will be contingent on the resources we have available to us; there might be ways in which we could change people's motivations, or economic circumstances, or institutions or cultural practices etc., but we may simply not have sufficient resources to do so. In such cases, the constraints look to be just as rigid and unalterable as hard constraints. This also suggests that it might be possible for a soft constraint to become a hard constraint over time, as the resources needed to change it are depleted.

In the previous chapter, I suggested that one simple way to categorize the various kinds of constraints is to say that constraints can be logical, physical or moral, and that within the category of physical constraints we ought to be particularly aware of the sub-categories of epistemic and psychological/motivational constraints.

Each of these simplified categories connects to a necessary condition for a political proposal to be both achievable and desirable. For a proposal to be achievable, it must be at least consistent with logical constraints. It must also be the case that we have sufficient physical resources to implement the plan, that we understand what the problem is and how to implement the solution in question, and that we are sufficiently motivated to use our knowledge and our resources to produce the outcome in question.

In addition to these conditions, which determine a proposal’s achievability, for a proposal to be desirable it must at least be the case that its implementation would not violate what we can call moral constraints. Recall that for a proposal to count as desirable, it must be the case that trying to pursue it will bring about some evaluatively desirable outcome, or satisfy some significant moral duty. Moral constraints rule out those proposals which would entail, for example, violating weightier duties or bringing about worse consequences overall – the content of these moral constraints will obviously be shaped by whatever broader moral theory one endorses. Assuming that we can have moral duties which are sensitive to the feasibility of our actions, making comparative feasibility assessments at the non-ideal level will be integral to the task of establishing whether there are moral as well as practical constraints that
prevent a proposal from being implemented, since feasibility amplifies the value of a potential political proposal provided that it is already desirable to some extent.

4.3 – Trade-offs and Fallibility

When all else is equal, feasibility considerations can determine whether we choose one political proposal over another. This is the most straightforward role feasibility considerations can play at the non-ideal level; given a set of proposals that are otherwise equally desirable, we always have reason to prefer the most feasible of the set, i.e. the one that stands the greatest chance of succeeding, provided that we try to implement it.

Of course, such cases are exceedingly rare, especially when it comes to the complex problems that non-ideal theorists tend to want to address. Some options will be more feasible than others, some will produce better outcomes than others, some will produce more egalitarian outcomes at the cost of personal liberty, and vice versa, for example. In cases where each option is similarly desirable, trade-offs will need to be made between the quality of an outcome and its probability of success.

At this point, we can draw on a distinction between different kinds of comparisons that will occur at the non-ideal level when deciding which political proposal to pursue.\(^\text{52}\) First, there are what we can call ‘route comparisons’, which involve comparing different political proposals that aim to bring about the same ultimate state of affairs. We make these comparisons in cases where we have a fixed goal in mind, but are unsure as to which proposals should be chosen in order to achieve the state of affairs in question (these comparisons may themselves include evaluative assessments of transitional states of affairs that lie between the status quo and the end-state). In contrast, ‘destination comparisons’ involve comparing different potential ultimate states of affairs to determine which is the most desirable. When we are dealing with a single variable, like feasibility for example, we can derive straightforward rankings of routes and destinations according to how feasible each one is.

\(^\text{52}\) The distinction between route and destination comparisons is also discussed by Gilabert and Lawford Smith (2012, p. 823): “This involves, in fact, a double moral consideration: the moral value of both the reforms’ results and the processes leading to them… Political judgement is always needed to determine the appropriate balance between process-related and result-related considerations.” On “hard cases” requiring trade-offs between different values, see Gilabert (2012) pp. 122-125, 248-248 and 263-266
However, making all-things-considered judgements involving proposals that vary across a number of values makes it much more difficult to select the proposal that we ought to pursue. We are forced not only to make route comparisons and destination comparisons, but to weigh routes against destinations in cases where the value of pursuing the best destination may be outweighed by the costs associated with the proposals that aim at that destination. In such cases, we must consider whether it is all-things-considered more desirable to aim for a next-best destination, given the routes available.

It is possible to conceptualise this problem in terms of moral constraints in the sense that we want to ensure that the political proposal we choose does not require violating any overriding moral principles such that they would make the proposal all-things-considered morally impermissible. In order to ensure this, we need a way to identify and accurately describe the normative considerations that ought to inform our decisions, when considering, for example, whether to aim for a highly desirable destination with little chance of success, or whether to choose a proposal that aims for a less desirable destination with a greater chance of success.

While we must take moral constraints into account at the ideal and non-ideal level, the process of identifying these constraints happens as part of a more abstract process. What I have in mind here is simply that whatever conclusions we reach about our preferred moral theory more generally are incorporated into our theories of justice such that we avoid producing recommendations which would achieve better states of affairs than the status quo, but which would be morally impermissible to pursue. In other words, we need to be aware of moral facts that will help us exclude from consideration political proposals that would achieve a valuable outcome via morally impermissible means. Political proposals excluded for these reasons will be achievable, but not desirable.

This is not to assume that there will only ever be one single option that we are required to choose. It may be the case that we are sometimes permitted to choose any of a range of available options. In such cases, however, we still need to identify the facts that define the set of permissible options. Either there are facts that tell us how to rank each available option, or there are facts that tell us how to define the set of permissible options, but unless we have access to these we risk violating moral constraints when we make trade-offs between values in selecting particular recommendations.
Thus far, I have been treating feasibility considerations in an objectivist sense. When we take an objectivist perspective on feasibility, we treat feasibility as though it is a property of a political proposal. This is the perspective we take when we say things like “this proposal has a 50% chance of success, but that proposal has a 90% chance of success”.

In contrast, a subjectivist perspective of feasibility focuses not on properties of the proposal itself, but on the properties of the agent making the feasibility assessment. Rather than reporting some fact about the proposal, subjective feasibility assessments report facts about the degree of confidence we have in our predictions that a proposal will succeed, if we attempt to implement it. These sorts of feasibility assessments can be seen as a response to epistemic constraints, both hard and soft. For the sake of simplicity, I shall focus on cases where there is broad agreement over these facts, and set aside the additional complications that arise in cases where there is disagreement about how to describe the relevant epistemic constraints.

When we make a feasibility assessment from a subjectivist perspective, what sort of claims are we making? Consider the sorts of constraints that might render a proposal impossible to implement. These include logical constraints as well as constraints based on unalterable facts about the laws of physics and human nature, but I have also suggested that they could include facts about human psychology, as well as epistemic, institutional and cultural facts, provided that we lack sufficient resources to alter such constraints. If we say that a proposal is feasible in the subjective sense, we say that we believe there are no hard constraints that would render the proposal impossible to achieve. If we say that a proposal is more feasible than another, in the subjectivist sense, we are making a claim about our beliefs about the relative strength of soft constraints that need to be tackled if we are to implement the proposal in question.

Such claims might be mistaken if there are constraints we have failed to recognise. While it might be possible to implement a proposal even if we have failed to recognise every constraint, subjective feasibility assessments will be more likely to align with objective feasibility to the extent that the proposals are produced by theorists who are aware of all possible constraints. Call this the completeness condition.
Even if we have successfully identified every relevant constraint, we might still make incorrect feasibility assessments in cases where we misidentify the status and the strength of some of those constraints. If we label a particular fact as a hard constraint when it is in fact a soft constraint, or vice versa, we will be less likely to make accurate and efficient predictions. If we label as hard a constraint that is in fact soft, we may rule out otherwise desirable proposals. Or, if we label as soft a constraint that is hard, we may fail to rule out straightforwardly infeasible options. We can also under- or overestimate the strength of particular soft constraints, by mistaking the extent to which they render a proposal more or less feasible. Subjectivist feasibility assessments will be more likely to align with objective feasibility to the extent that we have correctly labelled and assessed the strength of the constraints we identify. Call this the accuracy condition.

When we are concerned with achieving some particular goal, different constraints are likely to apply depending upon which route we take. In order to make an accurate feasibility assessment with regard to one particular proposal, which suggests one particular route, we need to successfully identify and accurately describe a sufficient number of constraints, the accommodation or removal of which will allow us to achieve our goal. A proposal will not necessarily be more likely to succeed than another merely because it succeeds in identifying a greater number of constraints, if it fails to identify a sufficient number to implement the proposal. Call this the sufficiency condition.

For example, suppose that a person is trapped beneath the bricks of a house that has collapsed following an earthquake, and we want to formulate a plan to rescue him. Some of the bricks will be too heavy for us to move; these represent hard constraints. Those that we can move come in various sizes; some can be removed very quickly while others take much longer to heave aside; these are soft constraints of varying strength. The completeness condition suggests that whatever plan we form will be more likely to succeed provided that we know about all the possible constraints; a proposal that fails this test might result in us tunnelling towards the victim only to find our path blocked by an immovable boulder, for example. The accuracy condition suggests that whatever plan we form will be more likely to succeed provided that we know which bricks are immovable and which are not, and provided that we can correctly predict how long it will take us to move each brick. Finally, the sufficiency condition suggests that our proposal must identify a path to the victim that targets a sufficient number of constraints; there may be multiple paths to the victim, but only one of these need be identified in order for the proposal to be more likely to succeed.
To the extent that any of these conditions are not met, or are only partially satisfied, it will be less likely that the proposal can be successfully defined and implemented. Failing to be aware of possible constraints will make it more likely that we fail to take into account some constraint that is a necessary part of the set of constraints we want to tackle. Failure to properly gauge the strengths of possible constraints will make it less likely that the proposal’s attempts to deal with constraints are successful, and failure to deal with a sufficient number of constraints will render the proposal straightforwardly infeasible.

Considering feasibility assessments from both the objectivist and subjectivist perspectives thus allows us to give a clear and comprehensive account of the problems that arise when we try to trade feasibility off against other values, and of the problems surrounding the reliability of our initial judgements about the values we are trying to trade off against one another.

Given that these are intractable problems, but given that we must produce action-guiding recommendations regardless, I shall now describe what I believe to be the best response to these obstacles. In effect, what follows amounts to a kind of “non-ideal” treatment of non-ideal theory, since it aims to describe how non-ideal theorists ought to proceed, given certain hard and soft constraints that make good, effective theorising more difficult to achieve.

Towards a “Non-ideal” Non-Ideal Theory

Each of the three conditions set out in the previous section is subject to potential fallibility in the sense that we may underestimate or overestimate the extent to which a proposal meets the three conditions. When we are assessing the evidence available to us, we depend on assumptions about the reliability of the evidence, the relevance of the evidence to the question at hand, our capacities to properly analyse the evidence, reach an accurate understanding of what our epistemic constraints are and what the evidence gives us warrant to predict in view of these constraints. The potential for error appears to be inescapably deep when dealing with the above sorts of questions. Each evidence-based claim about feasibility implies underlying assumptions about the confidence we have in our ability to make such claims; these claims imply underlying assumptions about our confidence to accurately gauge how confident we ought to be, and so on ad infinitum. We can call these “meta-feasibility claims” since they involve claims about the feasibility of making judgements about the feasibility of political proposals.
It is not possible for this chain to produce a finite regress, since there is no assessment that is not susceptible to the question “why are you so sure?”, and there is no answer to that question that does not result in a further assessment about the extent to which we can be confident in the preceding assessment. If this is a troubling implication, it should only be troubling to the extent that sceptical worries about the possibility of reliable knowledge are troubling more generally.

Additionally, examining our capacities requires resources that will eventually be exhausted provided that we go back far enough in the chain of assumptions. There will likely come a point at which we are faced with the decision as to whether or not we should act upon the information available, or continue to examine our epistemic constraints. It would surely be a mistake to think that the need to be sure of our assumptions always has priority over the need to act upon the information presently available.

While these sorts of deep sceptical worries should not trouble us, it is worth emphasising the role that epistemic constraints play in determining not just feasibility assessments, but practically every variable that goes into our all-things-considered decisions about what sort of political proposal we ought to pursue.

For example, Lisa Herzog has identified what she calls “the problem of knowledge” which she argues must be tackled by any ideal theorist who hopes to help guide theory at the non-ideal level (Herzog, 2012). The thought here is that if ideal theorists want to help create “concrete policy suggestions”, their views must eventually be able to incorporate complex non-normative sociological, psychological and historical facts about the real world. Herzog identifies several ways in which this knowledge might be inaccessible or unusable. For example, it may be too costly or complicated to attain, it may be too controversial or contested for it to be useful as part of the theory, or there may be no way to attain it without violating moral principles. When knowledge is unavailable due to a hard constraint, Herzog suggests this should cause us to revise any ideal principles that both depend on such knowledge and that aim to be action-guiding. Problems arise, however, once we consider cases where it is not clear whether the knowledge in question really is unattainable, or whether there are actions we might take now, to acquire that knowledge at some point in the future. This is an example of what I would call a “meta-feasibility” question about the reliability about our judgements about hard epistemic constraints.
Herzog’s solution to these kinds of problems is to emphasise “dynamic duties” to acquire knowledge necessary to implement principles of justice, bearing in mind the potential for feedback between the ideal and non-ideal. Discharging such duties will require philosophers to engage with experts in many other fields, in particular the social sciences, as they try to construct action-guiding recommendations for an increasingly complex world.

Epistemic constraints don’t just concern our ignorance of empirical knowledge that would help us identify possible political proposals and choose between them. We can also lack knowledge of, or sufficient confidence in, particular moral facts that would affect how we view the desirability of a political proposal under consideration.

If there are moral facts about what we ought to do when we don’t know what we ought to do, then the constraint-based view will label these as moral constraints that limit the political proposals we are permitted to choose under conditions of moral uncertainty, in addition to those constraints that apply under conditions of non-moral uncertainty.

Given these kinds of constraints, if we want to be able to make the right judgements about which political proposal to pursue, we will need to have the capacity to access to two kinds of facts. First, there are facts that will help us choose the most plausible principle for decision-making under conditions of moral uncertainty. Second, there are facts that will help us make better decisions, given some particular theory of decision-making under conditions of moral uncertainty.

To illustrate these different kinds of facts, consider the following brief descriptions of two views about how we should make decisions when we are uncertain about relevant moral facts. On one view, when we are faced with a choice between endorsing A, which is a moral view that says we ought to φ, or B, which is a rival view that says we ought not to φ, what we ought to do is figure out which option will give us the “highest expected objective value”. We figure out the expected objective value of A, for example, by assigning a value to φ on the assumption that A is true, and multiplying that value by our credence in A’s being true. An alternative view suggests we should compare the difference between φ and not-φ according to A with the difference between φ and not-φ according to B via recourse to some more general moral view which will provide us with a common metric by which to compare

---

53 This is one version of a view about decision-making under conditions of moral uncertainty considered by Andrew Sepielli (2009, pp. 10-12).
the values across both cases and to then multiply the resulting values according to our relative credences of A and B.\footnote{See Sepielli (2009). Both of these approaches require that we appeal to some more fundamental common metric as a basis of comparison and it should be noted that one might be sceptical about the existence of such a metric. The difference between the two views described above is that the first view begins by comparing the value of $\phi$ according to A with $\phi$ according to B whereas the second view begins by comparing the difference between $\phi$ and $\neg \phi$ according to A with the difference between $\phi$ and $\phi$ according to B.}

On either account, what we ought to do is affected by our credence in the options available to us. For any plausible view, facts affecting our credence in a moral claim will always be relevant when it comes to figuring out what we ought to do. Thus, if we can uncover facts that can help us adjust our credences in the relevant moral claims, this will help us to arrive at better decisions according to whichever decision-making theory we find most plausible.

In addition, there will be a distinct set of facts that will bear upon our confidence in the correct decision-making theory itself. The idea of improving our capacities to recognise relevant moral constraints covers both of the capacity to make accurate claims about which decision-making theory we ought to endorse, and the capacity to make accurate assessments about things like our credence in a particular moral claim, in order to ensure that the decision-making theory is more likely to produce the best recommendations when it comes to selecting political proposals.

So, to be sufficiently confident of our descriptions of the variables which inform our assessments of political proposals, including our understanding of moral constraints, and of the methods we use to combine the variables into an overall calculation of all-things-considered desirability, we need to know how feasible each of our claims are, given the constraints that apply to the theorists making the claims. Given that improving our ability to make accurate predictions may require resources that may in turn make our options less likely to succeed, it might be useful to rely upon certain kinds of heuristic principles when it comes to ensuring a sufficient degree of confidence, especially if we are dealing with extremely complex or controversial claims. This could yield a preference for assumptions that are less controversial over assumptions that are more controversial, or a preference for assumptions that are risk-averse over assumptions that are not, or even options that require fewer assumptions overall, and thus provide fewer opportunities to make mistakes.

\section*{4.4 – A ‘narrowing down’ approach}
Imagine a set containing all the possible options available to us to address some particular issue of injustice. Having ruled out the impossible, we must try to identify the most (normatively) desirable political proposal from the options which remain. It will rarely be the case that every remaining desirable option is equal across every normatively-significant variable except for one proposal, which has a higher desirability value because it is, for example, more feasible than any of the alternatives. Usually, we will find ourselves left with several options that differ across multiple variables.

A game of chess makes for a useful analogy to illustrate how we could go about narrowing down our options when we find ourselves unable to determine the optimal action to take. Chess has certain rules that constrain the number of possible moves a player may make at any given point in the game. These rules are analogous to hard constraints. Simply knowing what moves we could choose is not sufficient, obviously; we want to know whether there is a best move we could make, and if not, we want to know which moves would be better and which would be worse. Sometimes there will be a move that is clearly superior to all others. Sometimes we find ourselves with options that we know are equally as good as each other, any of which will be sufficient to win the game.56

Usually, however, it will not be clear precisely what the optimal move is. We assess the position of all the pieces, try to consider all of our possible moves, and then try to rule out the moves that appear to be obviously suboptimal. What we are left with will often be several options, with no way to be certain which of these is best. Sometimes this will be because there is no best move we could have made, sometimes (as in chess) this will be because there was a best move, but no way for us to know it and sometimes this will be because there was a best move but we have made an error of judgement in failing to recognize it as being the best.

The skill required to narrow down one’s options is certainly something that can be learned and developed in the context of chess. If the analogy holds in the context of selecting between different possible political proposals, we can ask what sorts of skills or information

---

56 One might object to the chess analogy on the grounds that all possible moves are equally feasible. While this is true, the analogy is intended to capture the contrast between the capacities of the novice and the grandmaster when it comes to identifying the subset of possible moves which will be more likely, if attempted, to lead to victory. In real-world cases, it is certainly true that each available proposal, in contrast to each possible chess move, will rarely be equally feasible. This should be seen as an additional complication which needs to be factored in to our ultimate all-things-considered judgements about which proposal to pursue, but the chess analogy allows us to consider a case where this complication does not apply, in order to focus on the capacities which need to be developed in order to pursue the narrowing down approach. I am grateful to an anonymous referee for bringing this point to my attention.
are necessary in order for us to be able to construct as narrow a set of options as possible from which to choose a political proposal. Even in cases where no further ranking is possible, we still have good reasons to prefer that the choices we ultimately select from are as narrow as we can make them since this increases the chances of choosing the best option.

We can draw some general conclusions from the above, as follows:

1. It is possible that not all feasible options may be ranked on a linear scale such that we can determine which one is optimal (either because there is no best option or because it’s impossible to know which is best).

2. Despite this, if there are better or worse options, we can increase our chances of choosing a better option if we are able to narrow down the set of good “moves”.

3. Because our judgements are always fallible, one important way of narrowing down our options will involve increasing our capacities to make reliable judgements about each component of an option.

On this view, feasibility considerations continue to play a role in ‘narrowing-down’ judgements, i.e. judgements where we exclude one option as not being part of the set of ‘good moves’. This approach could also be thought of in terms of ranking different types of options against each other, rather than trying to come up with a linear order of preference for each option. Suppose, for example, that we have nine possible moves we could make in some game of chess; three of which will leave us at a material disadvantage to our opponent and in a worse position, three of which require us to sacrifice pieces for positional advantage, and three of which will provide material advantage but at the cost of positional advantage.

Taking the exclusionary approach in this case might involve us saying something like the following: “I’m not sure whether it would be better to generate a better positional advantage, or to capture more of my opponent’s pieces. But I’m quite confident that either of those options would be preferable to moves where I can do neither. These three moves seem to give me neither kind of advantage, so I will exclude them from further consideration.”
Now consider the sorts of worries that are often expressed about ideal theory in general. We can make mistakes in specifying what ideal state of affairs we ought to aim for, or in specifying what constraints stand in the way of attaining the ideal, or in describing what measures need to be taken to successfully overcome these constraints. The possibility of miscalculations when it comes to feasibility in particular lies at the heart of some of the most persistent worries about ideal theory. A theory that assumes that its recommendations are feasible, when in fact they are not, is not merely a bad theory, but has the potential to be actually harmful if we attempt to follow its recommendations.

Given that each variable that determines the desirability of a political proposal is subject to potential fallibility, we can minimize the chance of making mistakes, by maximizing our capacities to identify and measure constraints.

The need to have sufficient capacities to gauge political proposals is not something we can avoid, given that inaction itself must be considered as just a different kind of proposal. Sometimes the best thing we can do will be to do nothing, but to know that that is in fact the best thing we can do will require the same sorts of capacities we need when deciding on political proposals in general.

4.5 Conclusion

I began this Chapter by suggesting that we should follow Gilabert and Lawford-Smith in thinking of feasibility as functioning in a binary role at the level of ideal theory, and in a scalar role at the level of non-ideal theory. My account builds upon Gilabert and Lawford-Smith’s view with the addition of moral constraints, the suggestion that any type of constraint can become a hard constraint, and the identification of the problems of trade-offs and fallibility, which need to be addressed if we hope to produce action-guiding recommendations that are achievable and desirable.

I then suggested three conditions which track the reliability of our judgements about trade-offs and about the feasibility of making accurate feasibility assessments. According to this view, we can have more confidence in our judgements to the extent that our theories succeed in identifying all relevant constraints (the completeness condition), in correctly labelling the status of constraints as being hard or soft and in correctly gauging the strengths of soft
constraints (the accuracy condition), and in proposing a plan that deals with a sufficient number of constraints in order to achieve the desired outcome (the sufficiency condition).

Once we understand the role that concerns about feasibility and fallibility should play in processes designed to produce action-guiding recommendations that are achievable and desirable, the importance of focusing upon our capacities to identify, assess and respond to moral and empirical constraints upon potential political proposals becomes clear.
5. Transitional Theory

5.1 Introduction

In this chapter, my aim is to investigate those principles that govern the transition from an existing non-ideal state of affairs to an ideal state of affairs, or to a ‘second best’ alternative.\(^{57}\) Such principles serve two general functions. First, if we begin with some particular normatively significant value or values in mind (such as ‘justice’ or ‘equality’ for example) and the assumption that we ought to try to reform existing political institutions until we reach the point at which these values or ideals are fully realized, transitional principles will allow us to articulate the necessary and sufficient conditions which must be met in order for us to deviate from these sorts of reforms. We can call this the ‘destination-selecting’ function of transitional principles in cases where the principles determine whether we ought to formulate proposals aimed at the ideal or at a second (or third, fourth, etc.) best state of affairs.

There will be two sorts of cases where deviation is permitted or required according to the correct transitional principles. First, there will be cases where we ought to maintain the status quo, due to the presence of normative reasons identified by the correct transitional principles. In other cases, transitional principles will provide us with recommendations that prescribe a transition to a new state of affairs which is more desirable than the status quo, but which does not move us towards the ideal (and which may make achievement of the ideal less likely). The second section of this chapter focuses on describing what the content of these principles might be, or in other words, what sorts of decisive reasons we can have in order to deviate from the pursuit of an ideal.

The third section of this chapter focuses upon the second general function of transitional principles, which is to help us to rank alternative political proposals aimed at achieving the same outcome in terms of their overall or all-things-considered desirability. Whereas the first function of transitional principles is concerned with selecting our destination, in terms of achieving some particular state of affairs, the second function can be described as a ‘route-

\(^{57}\) Obviously the alternatives we pursue in many cases will not be ideal or second-best, but third-best, fourth-best etc. etc. For the purposes of clarity and brevity, this chapter is restricted in focus to just the relationship between the ideal and the second best alternative(s). I assume that having worked out the circumstances under which we may deviate from the pursuit of an ideal, we will also be able to extrapolate from this the circumstances under which we should deviate from the pursuit of any less than ideal option in favour of a next best (even less ideal) alternative.
selecting’ function, given that it compares and attempts to rank alternative political proposals aimed at achieving the same outcome, via different means, or ‘routes’.

Making this distinction between principles which determine the appropriate destination and principles which determine the appropriate route is helpful because, as shall become clear by the conclusion of this chapter, different sorts of considerations apply depending on whether we are trying to decide if we ought to deviate from the pursuit of an ideal, or whether we are trying to decide which route we ought to take, given the state of affairs we intend to bring about. However, it is important to emphasize at this point that route-selecting and destination-selecting processes do not operate in isolation from one another. When selecting between different possible destinations, the available routes will be a relevant consideration. For example, if some desirable Destination A is preferable to desirable Destination B, but there are no desirable routes (or political proposals) which would allow us to arrive at Destination A, then this fact will be relevant, and usually decisive\textsuperscript{58}, in explaining why we ought to aim for Destination B (provided that a route to B is available), all else being equal.\textsuperscript{59} Similarly, the desirability of a political proposal will be determined in part by the state of affairs it would ultimately bring about if successfully implemented.

\textbf{5.2 Destination Selection}

One useful feature of the Constraint-Based View is that it provides a clear and helpful way to think about the transitional process between a non-ideal status quo, and the complete realization of some ideal state of affairs. Filtered through constraint-based language, the phrase ‘status quo’ refers to some particular state of affairs described in terms of all of the relevant hard and soft constraints with regard to some particular normatively significant value or values. A hard or soft constraint is ‘relevant’ in this sense, to the extent that it describes a fact about the world that is part of the explanation as to why the realization of the ideal state of affairs in question has yet to be achieved. For example, if the ideal state of affairs is a world without polio, a relevant soft constraint might be that not enough funding has yet been

\textsuperscript{58} It may sometimes be desirable to pursue a destination that we have no hope of achieving if doing so produces desirable consequences.

\textsuperscript{59} There may be routes towards A which are achievable, but not desirable, because, for example, they would require us to spend a significant amount of time in a state which is much worse than the status quo. Sometimes we must make things worse before we can make them better, but sometimes the long-term benefits will not be worth the short or medium-term costs.
invested in vaccination programs (this constraint might itself be explained by further soft constraints, such as epistemic constraints about how to generate sufficient funding, or motivational constraints which prevent people from being sufficiently motivated to generate and invest the necessary funding, etc.). A relevant hard constraint might be that human beings are susceptible to disease. A complete description of the status quo in this sort of case would include all the facts that might stand in the way of our ability to achieve a world without polio. (In practice, of course, the proposals that we produce and intend to implement in the real world can be effective even if they do not include every single relevant fact when taking account of the status quo. As indicated in the previous chapter, it is important is that a proposal addresses a sufficient number of constraints in order to achieve its goal.)

Maintaining the constraint-based perspective, we can recall a method sketched in previous chapters that describe how we might produce political proposals, i.e. proposals intended for the real world that aim to move us towards the ideal. First, we identify the value(s) we wish to implement and then we describe an ideal state of affairs. To describe the ideal state of affairs with respect to the value in question is to describe a world where the value is fully realized given all relevant hard constraints. The next step is to give an account of the status quo, as described above, before comparing each description in order to identify some set of soft constraints, the removal of which will be sufficient to fully realize the ideal. What emerges from this process is a political proposal, which will be a plan of action that aims to remove one or more soft constraints that stand in the way of the ideal. A proposal may address a full set of soft constraints such that the successful implementation of the proposal will result in the ideal state of affairs being achieved. Or a proposal may target one or more constraints whose elimination is necessary but not sufficient to achieve the ideal, as part of some wider set of constraints that must all be removed in order for the ideal to be achieved. To remove or eliminate a constraint in this context means to perform actions that have the effect of making some fact(s) no longer true, or making it such that true facts no longer figure in an explanation as to why the ideal cannot be achieved. For example, if a proposal suggested doubling the investment into polio vaccination programs, it could then be false that

---

60 This identification process is not part of the application of the Constraint-Based View itself – we can identify important values or combinations of values via the theory of ideals or via some other antecedent theoretical process. These can include prior applications of the constraint-based view to other values, if the conclusions of those applications are expressed in terms of normative principles which are then built into a subsequent application of the constraint-based view. For example, if our theory of ideals tells us that liberty and equality are both valuable, but that liberty must always have priority, we might produce a constraint-based view of liberty, before turning to a constraint-based view of equality where the value of equality is constrained by whatever we have already determined is necessary to protect the value of liberty.
insufficient resources have been expended in order to achieve a world without polio. Or, if an alternative proposal suggested some extremely efficient way of delivering vaccination programs (making additional investment unnecessary), it would then be no longer true that the fact that only X amount of money had been invested in the vaccination programs figures in the explanation as to why the ideal has not been achieved.

The first function of transitional principles is to specify the conditions under which these sorts of political proposals that aim to take us closer to the ideal may be suspended or abandoned in favour of alternate proposals that may suggest that we maintain the status quo, but will more often suggest that we aim to bring about some new state of affairs which, while preferable to the status quo, may make the ideal more difficult to achieve. These alternate or ‘second best’ proposals will have the same sort of structure as those aimed at moving us towards the ideal. The only substantial difference is that second-best proposals will be directed at realising non-ideal states of affairs.

The fact these sorts of proposals offer second best states of affairs suggests that if we want to deviate from the pursuit of the ideal, we ought to be able to provide strong reasons that justify this deviation, especially if our second-best proposals would have the effect of hindering our ability to pursue the ideal. The transitional principles that govern this kind of deviation should be thought of as a distinct part of non-ideal theory more generally. According to the constraint-based view, non-ideal theories are theories that will tell us what we ought to aim at, given certain hard and soft constraints. Some of these (hard) constraints may be normative constraints that prohibit us from pursuing the ideal. The aim of transitional principles is to articulate this set of potential normative constraints that can require or at least permit deviation from pursuit of the ideal. The part of non-ideal theories that transitional principles affect is that part which is sensitive to these particular ‘deviation-permitting’ or ‘deviation requiring’ constraints. Depending on whether and how such constraints apply in any particular case, non-ideal theories more generally may or may not have achievement of the ideal as their ultimate aim. So, what do these constraints look like? What are the facts that, if true, would permit theorists to formulate proposals that might have the effect of making the ideal more difficult or even impossible to pursue?

**Physical Constraints**

According to the constraint-based view, when we describe the ideal state of affairs with regard to some significant normative value(s), our description ought to include a relevant set
of hard constraints. These are facts about the world which we have strong reasons to think are unalterable, and they are relevant to the extent that their truth has implications for our ability to implement the normative ideal. Such facts will almost always be assumed implicitly in a description of the ideal, rather than explicitly enumerated, and will generally be fairly easy to identify.

Given that a description of the ideal state of affairs will, by definition, include the presence of hard constraints, the ideal state of affairs will necessarily be achievable in principle, provided that our description of the ideal state of affairs is completely accurate. The most straightforward case where we might think our obligations to pursue an ideal no longer hold, would be cases where that ideal is no longer achievable, provided that we accept the idea that ‘ought’ implies ‘can’.

The constraint-based view suggests several ways in which an ideal may become unachievable. First, an ideal can ‘become’ unachievable in the sense that we discover that our initial description of the ideal state of affairs was mistaken. Perhaps there was some fact, which we assumed to be a soft constraint, but turns out to be a hard constraint, or perhaps there was some hard constraint of which we were unaware, the presence of which constrains us more than we initially realised. In this sense, the ideal was never truly achievable in the first place. For example, suppose that state officials have good reason to believe their state will generate a large amount of money following the discovery of a certain quantity of offshore oil, only to discover at a later point that the initial estimations as to the quantity of the oil were incorrect, and that the expected oil-based revenue is now much lower than originally anticipated. Following the discovery of the error, the officials need no longer pursue an ideal that relied upon the assumption that they would have access to a large amount of money generated via revenue from oil.

In addition to these sorts of cases, an ideal may become unachievable when we create additional hard constraints, either by pursuing political proposals aimed at some alternative destination, by failing in our efforts to pursue political proposals aimed at the ideal, or by expending resources more generally. To modify the example from the previous paragraph, suppose that a state discovers large amounts of oil and intends to use the revenue to bring about an ideal state of affairs where every citizen has access to a certain level of education and healthcare. Instead of the money being used for this purpose, however, suppose that the state becomes engaged in a war and uses the money to attack its neighbour, or defend itself
from attack. Depending on the nature of the war, diverting the resources may or may not be required (I shall return to this idea below), but in either case the result is that the previously achievable ideal becomes impossible to achieve due to the presence of new hard constraints which track the lack of sufficient resources caused by the decision to use the revenue for the war effort instead.

At this point, it is worth making a distinction between temporarily impossible and permanently impossible ideals. In cases where we have good reason to believe that an ideal is permanently impossible, we will usually have decisive reasons to deviate from the ideal (unless we have reason to believe that the pursuit of the ideal is likely to bring about the best available state of affairs). If we believe, however, that the ideal may be achievable at some point in the future, then the fact that it is presently unachievable is not likely to be sufficient to give us decisive reasons to pursue an alternative destination. In such cases, we may have what Mark Jensen (2009) refers to as a diachronic ability to take actions now that will have the effect of rendering an ideal achievable at some point in the future. We can describe the exercise of these diachronic abilities in constraint-based terms by imagining political proposals aimed at identifying and eliminating the constraints that make the ideal temporarily impossible to achieve. Given all of the above, we could formulate the first transitional principle as follows:

**TP1 (Impossibility):** If we have sufficient reason to believe that a particular ideal is not achievable, both now and at any point in the future, and if ought implies can, we cannot be obliged to pursue that ideal and it is therefore at least permissible to deviate from the pursuit of that ideal.

This principle may seem so obvious that it requires no further defence, but we could imagine a case whereby the pursuit of an ideal, even an unachievable one, is considered valuable because it serves to produce certain valuable results regardless of whether or not the ideal is actually achieved. A devout Christian, for example, may see the ideal as a world where she perfectly emulates the moral character of Jesus, and her efforts to do so may turn out to make her a more virtuous person, as well as making things go better for other people in the world, even if she never succeeds in her ultimate aims. Even if a goal is impossible to achieve, it looks as though the pursuit of some goals can be valuable regardless, and this may be viewed as an objection to TP1, above.
One response to this sort of objection might be to reply that TP1 merely licences departure from the ideal, rather than demanding that we seek a second-best solution. This reply is probably too weak, however, given that we could imagine cases where pursuit of an unattainable ideal is more valuable than any alternative. In some such cases, it looks as though we might be actually required, rather than merely permitted, to continue the pursuit of an unattainable ideal.

A better response to such cases would be to suggest that the ideal be redescribed to take into account the value created by the pursuit of an actually unattainable state of affairs. To take the case of the Christian who tries to achieve an unachievable moral character, we could describe the ideal state of affairs as one where the Christian consistently does “the best that she can do” when it comes to making moral decisions, for example. It could be the case that the most reliable way to ensure that the Christian behaves consistently in this way is for the Christian to act as though the ideal of behaving exactly as Jesus would have behaved is achievable, even though it is actually impossible.

This suggests an additional hurdle that must be overcome before we can use TP1 to justify deviation from the pursuit of the ideal. It is necessary to show, not merely that the ideal is unattainable, but also that the pursuit of the unattainable ideal is not more valuable than any alternative proposal under consideration. So, we can articulate the following revised version of the principle:

**TP1* (Impossibility):** If we have sufficient reason to believe that a particular ideal is not achievable, both now and at any point in the future, and if ought implies can, and if the pursuit of the ideal is not more valuable than any alternative proposal under consideration, we cannot be obliged to pursue that ideal and it is therefore at least permissible to deviate from the pursuit of that ideal.

More generally, it might be objected that the way I have been discussing ideals makes it sound as though ideals are always states of affairs to be achieved, and that this commits me to an overly consequentialist perspective. But we might think many ideals are standards of conduct or ways of regulating our behaviour that apply differently in different contexts, such as the “ideal” of acting in ways you can justify to those whom your actions affect, for
example. It is not obvious, so the objection might go, how an ideal like this fits into the conceptual framework I have been describing.\textsuperscript{61}

There is a view about moral theories in general, which holds that “for any moral theory M, there is some conceivable theory of the good that, when combined with the consequentialist principle “φ-ing is morally permissible iff φ-ing would produce the best available state of affairs,” yields moral verdicts that are, in every instance, identical to those of M.”\textsuperscript{62} Roughly speaking, the basic idea is that any moral theory can be described in consequentialist terms. For example, suppose we endorse a moral principle, which holds that it is never permissible to, lie regardless of what the consequences of lying happen to be. At first pass, this looks like a decidedly non-consequentialist view, but it would be possible to “consequentialize” this view by defining “the best possible state of affairs” as “a state of affairs where people never lie”.

I do not have the space to defend this view here, but it is a view that I endorse. It seems clear to me that we can talk about ideals as “states of affairs” in a way that can accommodate ideals concerning standards of conduct or behaviour. In such cases, the ideal state of affairs should be considered one in which the relevant standard of conduct is fully realised. This is compatible with the idea that such standards might apply differently in different contexts – so long as the ideal state of affairs is understood as any state where the relevant moral principles are being applied faithfully and consistently.

Finally, note the requirement in TP* that we have sufficient reason to make the various claims we need to make about the pursuit of the ideal in question. The sufficient reason requirement can be derived from a more general moral requirement which we might think of in terms of the notion of “due diligence” when it comes to formulating political proposals in general. To return to the example of the state that strikes oil, suppose that the state officials in charge of the national budget are negligent when it comes to paying for extensive surveys of the newly-discovered oil fields, and that they base their budget on initial estimations that are known to be highly unreliable. In Scenario A, suppose that the initial survey has greatly underestimated the amount of oil present, and that the state officials decide to pursue a second-best solution which is in fact much less desirable than the state of affairs which would have been brought about had they known the truth about their resources. In Scenario B,

\textsuperscript{61} I am grateful to Jonathan Quong for making me aware of this objection.

suppose that the initial survey has greatly overestimated the amount of oil present, and that the state officials decide to pursue an ideal that is in fact impossible. In Scenario A, the state officials lack sufficient reason to believe that the ideal is not achievable, and we can presume that people are likely suffer avoidable harms as a result of their negligence. In Scenario B, an unachievable ideal is pursued though the state officials lacked sufficient reason to ensure that it was achievable before pursuing it. Assuming that the pursuit of the unachievable ideal is less preferable to a budget based on a more accurate assessment of the expected revenue generated from the oil, the officials will likely cause avoidable harm to citizens due to a highly inefficient and wasteful use of resources.

Following the above, we can formulate a second transitional principle:

**TP2 (Due Diligence):** In order to deviate from or to pursue an ideal, all reasonable efforts must be made to ensure that the decision is justified. This entails taking measures to ensure that we have sufficient capacities to ascertain whether a proposal is achievable, to what extent it is desirable and that we have sufficient capacities to apply transitional principles themselves.

As with the previous principle, TP2 may seem quite straightforward and uncontroversial, give that it merely expresses the claim that we must not expose people to unnecessary risks. In practice, however, much will depend on precisely where we set the bar when determining whether the due diligence requirement has been met. For example, it seems reasonable to suggest that the extent to which a proposal should be regarded as risky is sensitive to the level of confidence we have in our ability to make feasibility assessments. If the bar is set too high in terms of the capacities we require in order to be sufficiently confident that a proposal will succeed, the risk is that we exclude viable options for the sake of greater epistemic certainty. On the other hand, setting the bar too low risks including options which expose people to impermissible levels of risk.

It may be objected that TP2 is vulnerable to an infinite regress of the following sort: if satisfying the due diligence requirement requires taking sufficient measures to ensure that we have a particular set of capacities, is there also a requirement to take sufficient measures to ensure that we have the capacity to take sufficient measures to ensure that was have a particular set of capacities? Despite this worry, we should not think that endorsing TP2 commits us to an infinite regress of capacities to ensure capacities. Given that developing capacities requires resources, there will likely come a point in the chain where further
investigation or development of our capacities would be likely to negatively affect our existing capacities to pursue what seem to be achievable and desirable political proposals. When we reach such a point, the regress can be halted, given that we will have reasons to stop focusing on the capacities that we might develop, and start acting on those we possess.

**Demandingness Constraints**

Suppose that we wish to deviate from the pursuit of an ideal not because it would be impossible for us to achieve it if we tried, but because the only route(s) available towards that ideal would entail that we bear burdens that are overly demanding. Should demandingness play a role in transitional principles such that it is sufficient to deviate from pursuit of the ideal to demonstrate that such a pursuit would be overly demanding, and if so, how should we understand the threshold beyond which pursuit of the ideal should be considered overly demanding?

First, we should note that a proposal’s being overly demanding does not necessarily result in an obligation that we deviate from pursuit of the ideal. That a proposal is overly-demanding will not be sufficient to require deviation (rather than merely permit it) if it is the case that what makes the proposal overly-demanding is that it requires people to voluntarily take on burdens which we believe they have no duty to bear, according to our preferred theory of morality. In such cases, the fact that a proposal is overly-demanding will merely permit departure on the basis that pursuit of the ideal is supererogatory.

We should also note that there can be situations where we might think that no burden is too great to bear, but the clearest examples of such cases typically involve people who have only rendered themselves liable to bear such heavy burdens due to some prior forfeiture of rights. For example, if Alice deliberately attacks and injures Bob such that the only way for Bob to survive is for Alice to donate a vital organ to Bob, then it looks as though Alice cannot appeal to the fact that donating her organ would be overly demanding as a way to avoid her obligation towards Bob, even if the cost of the donation is Alice’s own life.\(^6\)

---

\(^6\) One may not share my intuition that Alice would be morally required to donate the organ to Bob if one believes that certain rights to bodily integrity cannot be forfeited under any circumstances. However, while it may be the case that nobody has the right to coerce Alice into donating her organ, we may still hold that she is morally required to do so.
To give the case for a demandingness constraint a fair chance, therefore, I shall only consider those cases where all agents involved are not guilty of any prior moral wrongdoing and have not forfeited any rights they may have had against bearing overly demanding burdens.

One way to ground a demandingness constraint would be to adopt a strategy similar to that taken by Bob Corbett in responding to arguments about the obligations we have towards complete strangers made by Peter Singer and James Rachels. Singer and Rachels begin with the principle that “if it is in our power to prevent something very bad happening, without thereby sacrificing anything of comparable moral significance, we ought to do it” and arrive at the conclusion that we bear deep and pervasive moral obligations of “beneficence” to strangers, such that we must attend to the necessities of any and all strangers before we may permissibly enjoy any luxuries ourselves. This sort of view may be considered a paradigmatic case where demandingness constraints should apply (assuming that they should apply at all).

Corbett’s counterargument to this view depends in part upon the idea that, if ought implies can, and if it would be simply too psychologically demanding for most human beings to prioritise the necessities of strangers over their own luxuries, it is not true that it is really “in our power” to meet the kinds of obligations Singer and Rachels think we have. If this is correct, it seems to follow that since it is not really in our power to meet these obligations, if ought implies can, there are no such obligations.

I say that “there are no” obligations above rather than “there can be no” obligations, because the above strategy does not distinguish between cases where we lack the ability to meet putative moral obligations due to some permanent unalterable feature of our psychological makeup, and cases where it might be possible for us to alter our psychological preferences or to subvert them somehow.

Consider, for example, the case of Victor, a particularly weak-willed vegetarian, who believes that he is morally obliged to refrain from eating meat. Despite this belief, Victor finds it extremely difficult to stick to a vegetarian diet. Nevertheless, Victor tries to develop

---

65 We should understand Corbett’s claim that “it is not in the power” of most people to prioritise the needs of strangers in this way as being a claim about the motivational constraints that most people experience when they try to act as someone like Singer would have them act. Understood in this way, the best objection to Corbett’s view is the empirical evidence which shows that at least some human beings really are capable of motivating themselves to act in these very demanding ways. How sympathetic we are to Corbett’s view should therefore depend on how pessimistic we are about the capacities of human beings in general to motivate themselves to comply with more demanding moral requirements.
new eating habits; he experiments with different kinds of vegetarian meals until he is able to cook food that he finds truly satisfying and manages to stick to a vegetarian diet. However, Victor’s daughter Victoria, who is also a committed vegetarian (at least in theory), is even fussier than Victor when it comes to food and finds herself unable to adjust to her new diet, even given Victor’s newfound culinary skills. In response to her dilemma, Victoria decides that since it is Victor’s job to do the shopping, she will give him instructions not to buy any meat-based products. As a result, Victoria finds herself forced to eat a vegetarian diet. Perhaps this strategy too might fail, leading Victoria to pursue a second-best solution such that she only eats meat every second day, but it seems clear that she ought to at least try to alter or subvert her preferences before abandoning the ideal state of affairs where she consumes no meat at all.

So, it is not sufficient merely to show that some putative obligation is so demanding that we are not able to meet it at present, owing to our psychological makeup. It must also be the case either that (1) there are no morally permissible non-supererogatory actions we could do to change our psychological preferences and/or (2) that there are no morally permissible non-supererogatory actions we might take to subvert our psychological preferences. In other words, if we find that pursuit of the ideal is too demanding, we need to know whether the psychological constraints that make it too demanding are hard constraints or soft constraints. If they are merely soft constraints, a further argument is required to show that we possess a right to refuse to alter these soft constraints, even at the expense of achieving the ideal. Here we might invoke the idea of “personal prerogatives” or “agent-relative reasons” to explain why agents might be permitted to deviate from the ideal.66 While the two notions are not identical, the general idea here would be that there is a protected sphere of action, or a certain class of reasons that explain why agents may give additional weight to the pursuit of their own life projects, or their own well-being, over the interests of others.

Given all of the above, we may adopt the following transitional principle:

TP3 (Demandingness): It is permissible (but not required) to deviate from the pursuit of an ideal if it is the case that the pursuit of the ideal can be reasonably expected to impose costs upon agents which are overly demanding, such that pursuit of the ideal would ask agents to bear costs which are greater than we believe the correct theory of morality requires.

66 For more on the idea of personal prerogatives see Scheffler (1993).
‘Deep’ Moral Constraints

Suppose we have a case where none of the previous principles can provide a reason for us to deviate from the pursuit of the ideal. In such a case, the ideal is physically possible for us to achieve, satisfying TP*, the pursuit of the ideal is not overly demanding, satisfying TP3, and the due diligence obligation (TP2) has also been satisfied. Suppose that the ideal we have in mind (owing to concerns about pollution and a sustainable global environment) is a world where the total population is below a certain level by the year 2020, and that one political proposal under consideration recommends that in order to achieve this we ought to exterminate 20% of the world’s population. In such a case, we can justify deviation from the ideal via an appeal to what I shall call ‘deep’ moral constraints. My aim here is to be neutral with regard to a particular brand of moral theory; it is sufficient to show merely that some underlying normative principles will apply, at least on any plausible moral theory. The only exception that occurs to me would involve a kind of consequentialist view that holds that the value generated by achieving the ideal is so great that it would outweigh any and all costs incurred along the way.

A more plausible moral theory might respond to the above case by invoking a deep moral constraint based on the idea that we must not use some people merely as a means for the benefit of others. Deviation from the ideal would be required on this view, if the political proposal has the effect of instrumentalizing some people merely so that others can benefit when the ideal state of affairs is achieved. A rule consequentialist view, on the other hand, may permit agents to be used in this way, but would require deviation from the ideal in cases where the only way to achieve the ideal requires behaving according to rules which would not produce the best consequences overall. A transitional principle that is sensitive to the normative constraints of a broader moral theory will have the following form:

**TP4 (Deep Moral Constraints):** Deviation from the pursuit of the ideal is permitted and may be required if the pursuit of the ideal would require the violation of some deep moral constraint, where this is understood as a fundamental moral commitment, the content of which is determined by the broader moral theory we believe to be correct.
By saying that a moral commitment is fundamental in this way, I mean that it is simply not ‘up for grabs’ or open to revision in the light of the question presently under consideration (i.e. ‘should we deviate from the pursuit of this particular ideal?’). Such commitments represent fixed assumptions about moral principles within the context of the present question. This is not to say that such assumptions must be fixed in any context, simply that they represent whatever broad, basic and foundational assumptions about morality that we think we are entitled to take for granted while thinking about whether we ought to pursue a particular ideal.

To use an analogy with epistemic assumptions, suppose we want to build a bridge so we are considering various different proposals as to how best to construct it. Within the context of this project (i.e. selecting the best proposal) there are a range of basic epistemic assumptions which we take for granted – we assume that the world is roughly as it appears to us and that we are not living in a dream or a simulation, that the laws of physics will behave consistently over time, and so on. In other contexts, we might have cause to challenge these assumptions, but it would be silly to object to the proposals of an engineer because she had not taken seriously the possibility that the world is a dream or that the laws of physics might be different tomorrow. Similarly, if there are a set of assumptions about moral facts which we think it is appropriate to take for granted, and which it would be silly to question in the light of the present project, these will constitute ‘deep moral constraints’.

**Publicity Constraints**

One might think that TP4 will be problematic given that there is reasonable disagreement as to what the correct moral theory happens to be. Suppose, for example, that we have some particular political proposal which does not violate any deep moral constraints from a deontological perspective, but which would entail the violation of deep moral constraints from a consequentialist perspective, or vice versa. Or suppose that we believe deviation from the ideal is required because of a deep moral constraint that forbids actions that transgress the will of God. If our transitional principles are too sensitive to controversial moral theories, especially theories which can be reasonably rejected, the worry is that the deviation from the ideal, even if it is justified on a purely objective sense, might not be capable of being justified to the actual people who must be coerced or suffer burdens in order for the proposal to be implemented. If this concern with public justification of deviation represents a moral
constraint that we ought to take into account, then we should include the following in our list of transitional principles:

TP5 (Publicity Constraint): Deviation from the pursuit of the ideal is permissible provided that the reasons for the deviation are ones to which all reasonable people could assent. These reasons may include the fact that pursuit of the ideal itself cannot be defended upon grounds to which no reasonable person could reject.

Applying this kind of principle in practice requires, among other things, a convincing story as to why public justification of this sort matters in the first place as well as a plausible account of what constitutes a principle being ‘reasonably rejectable’. For my present purposes, I will set aside the complications arising from how we might implement such a principle in practice and focus instead on the more fundamental question of why we should think that some sort of publicity constraint should apply in the first place.

We have two distinct reasons for caring about publicity in this context. The first is a purely instrumental reason – we have reason to care about publicity for instrumental reasons when we have reason to believe that abiding by a publicity constraint will lead to better outcomes in terms of our ability to realise valuable ideals. For example, it seems plausible to suggest that if we pursue only those proposals that cannot be reasonably rejected and if people tend to understand that the proposals being pursued cannot be reasonably rejected, people will find it easier to motivate themselves to act, and the states of affairs that people produce will tend to be more stable. Of course, these assumptions depend on other assumptions about what we might call the ‘communicability’ of the public reasons that we appeal to. That is to say, abiding by publicity constraints will positively affect people’s motivational constraints, and the stability of the states of affairs that they achieve, to the extent that people are actually able to appreciate and endorse these reasons. If, for example, we live in a society with large numbers of unreasonable people, or where reasonable people find it difficult to appreciate the reasonableness of political proposals, this will lessen the instrumental value of a publicity constraint.

The second reason we have for caring about publicity stems from a concern about the moral permissibility of coercing people to abide by laws that they could reasonably reject. This is a concern about the legitimacy of coercive political institutions, and one that we might derive from more fundamental moral constraints about respecting other human beings. On one version of this view, our moral obligations to respect other human beings require that we
are appropriately sensitive to the fact of ‘reasonable pluralism’: the fact that reasonable human beings can come to hold conflicting ideas about the good life as a result of what Rawls calls ‘the burdens of judgment’ (Rawls, 2005, p.58). These are facts that explain why deliberation among reasonable people can sometimes lead to disagreement through no fault of the parties involved, for example, due to disagreements over how to interpret empirical data or concepts, or over which substantive moral theory to endorse, and so on.\(^{67}\)

It might be objected that including TP5 here is unnecessary, in the sense that it should be subordinate to TP4 (deep moral constraints). If we care about publicity, so the objection might go, it is only because we care about some deeper moral obligation(s) to do with respecting people and the legitimate use of coercion. Since TP5 seems to be derivable from TP4, why should it merit its own category?

In response to this, we should concede that TP5 can indeed be derived from TP4, but note that TP5 also applies to TP4, and indeed to every transitional principle that I will set out in this chapter. Deviating from the pursuit of an ideal requires us to justify this deviation by offering an explanation as to how it is permitted or required by transitional principles. If publicity matters, therefore, then so does the publicity of transitional principles themselves.

This leads us to a second, related objection: isn’t the idea that publicity matters itself highly controversial? If so, it seems as though the publicity requirement cannot pass its own test.

There are several ways one might respond to this objection. One could dispute the claim that the idea that publicity matters is controversial, though the risks of this approach are that one ends up with an account of our values which is either so narrow that it is obviously controversial, or so broad and vague that it is difficult to see how a publicity requirement can be straightforwardly derived from it.

Or one could argue that the value of publicity, while controversial in the sense that people might support it for different comprehensive reasons, can be ‘converged upon’ by people who share different values but ones that each provides a justification for a publicity principle.

Or one could argue that what we might call the publicity principle does not need to apply to itself, because it is a kind of “meta” principle “that applies to about the appropriate

\(^{67}\) See Quong (2013).
conditions for moral or political justification, and not a first order normative claim within that discourse” (Quong, 2013).

Of these possible strategies, it is not clear to me whether attempts to show that a publicity principle is not controversial, or could be the subject of convergence are plausible. Regardless of the answers to those questions, however, we ought to believe that the publicity principle does not apply in fact apply to itself.

Part of Gaus’ strategy when addressing this question is to compare the publicity principle to a principle like that of falsification in the context of the scientific method. Asking that a publicity principle apply to itself is, so the thought goes, like objecting to the claim “a good theory is falsifiable” on the basis that that theory is not falsifiable. We don’t think that’s a good objection, so why take it seriously when a similar one is levelled at the publicity principle?

David Enoch notes, however, that some principles do apply to themselves and can only be made not to apply to themselves via ad hoc stipulations (Enoch, 2013). The challenge for those of us who think that the publicity principle doesn’t apply to itself is to show why such a stipulation is not simply an ad hoc way of avoiding the problem.

I think the best way to answer this challenge is to begin by asking the questions “why do people think a publicity principle is important” and “what is a publicity principle supposed to do?”

Public reason liberals think a publicity principle is important, because they value political legitimacy. This is the prior, more fundamental value that motivates public reason liberals to articulate the publicity principle as an expression of how the value of political legitimacy is to be promoted. The publicity principle is supposed to help secure political legitimacy by serving what we can call a “reason-selecting” function – when deciding which reasons we may use to exercise coercive power over citizens, the publicity principle selects some of these reasons (i.e. those which satisfy the publicity condition) as being potentially acceptable, thereby excluding the remain, non-public reasons from consideration.

See also Gaus (2011) pp. 227-228.

Of course, we might want to tell a longer version of this story, given that one may only care about something like political legitimacy insofar as one cares about something like “respecting one’s fellow human beings” and so on. Regardless of how far back in the story we go, it is sufficient for my present purposes simply to note that the publicity principle will be derived from some prior value.
Now, (within the context of choosing reasons to appeal to when justifying coercion to people) let us distinguish between what we can call a “reason-selecting principle” that tells us which reasons to select in order to promote some particular normative value (of which the publicity principle is one), and a principle that tells us which reason-selecting principle to select. Call these “principle-selecting principles”.

If the publicity principle applies to itself, and defeats itself, then it must serve not merely a “reason-selecting function” but also a “principle-selecting” function. In other words, reason-selecting principles like the publicity principle must be a “legitimate target” for the publicity principle, in order for its critics to turn the principle on itself.

Presumably, the publicity principle cannot have selected itself – that would be to say that the publicity principle is justified because the publicity principle is justified. Yet once selected via some higher-order principle-selection process, why should we not think that the question of principle-selection has been settled? We have considered our values and picked a reason-selecting principle, and now that reason-selecting principle will tell us which reasons to select when attempting to provide justifications to people.

For the critics’ argument to work here, we need to imagine someone being asked to accept a particular reason (or to reject a particular reason) on the basis of whether it satisfies the publicity requirement. When this person asks, “Why should I not appeal to this reason?” for example, the response of the public reason liberal is going to be something like “because the publicity principle is true”. Yet now it looks like we are giving the person a reason that (let us suppose) they might reasonably reject. Is this not precisely the sort of reason that the publicity principle says we cannot appeal to?

Actually, it isn’t. The critics’ argument here depends on a conflation between two different kinds of reasons. These are what I will call “coercion-justifying reasons” on the one hand, and “reason-justifying reasons” on the other. To see the distinction between the two, consider the following case. Suppose Bob wants to assault his neighbour, but the state uses coercion to prevent him from doing so. Bob demands an explanation. The direct explanation for the state’s coercion is going to be something like “we are entitled to coerce you, because your neighbour has a right not to be assaulted”. Here the state has given Bob a “coercion-
justifying reason” which we can define as any reason given in an attempt to justify coercion to a person subjected to it.\textsuperscript{70}

Crucially, “because the publicity principle is true” is \textit{not} a reason to coerce Bob. Rather, it is \textit{a reason to appeal to a particular coercing-justifying reason} (in this case “people have a right not to be assaulted”) which \textit{is} a reason to coerce Bob. In order to get us to answer, “Because the publicity principle is true”, Bob will have to ask not “why are you coercing me?” but rather “why are you giving me this reason for coercing me. A “reason-justifying reason” can be defined as a reason that is offered to a person in order to justify selecting some other reason to offer to that person.

This then, allows us to show why the publicity principle does not apply to itself – it is a principle for selecting coercion-justifying reasons. To ask someone to justify coercion by justifying the publicity principle is a mistake, therefore: it is to ask for a reasons-justifying reason in place of a coercion-justifying reason. According to the publicity principle, only the latter need be subject to a publicity constraint.\textsuperscript{71}

\section*{5.3 Route Selection}

Having considered the principles that determine the necessary and sufficient conditions under which deviation from the pursuit of the ideal is permitted, I shall now offer an account of the principles that determine which political proposals we ought to choose, given a particular (ideal or second-best) destination. My account shall proceed in two stages. First, I shall offer a list of the normatively significant factors that must be considered when assessing which political proposal we ought to attempt to implement. These factors (with one exception to be discussed below) apply regardless of whether we are aiming at an ideal or at a second best solution. Having given an account of the relevant features of a political proposal that determine its overall desirability relative to alternative proposals aimed at the same

\textsuperscript{70} Included in this definition then, are reasons which do not actually justify the coercion in question provided that they are offered with the aim of justifying the coercion.

\textsuperscript{71} One might object that if we really care about political legitimacy, why not care that “reasons-justifying reasons” are not also subjected to a publicity constraint? My tentative answer to this question is to say that the reasons we have for caring about coercion-justifying reasons are not the same as those we have for caring about reasons-justifying reasons. We care about the former because we don’t want reasonable people to be coerced for reasons that they cannot endorse. But we care about the latter because we think that political legitimacy is important. It is from this value that the publicity principle is initially derived, so to demand that it be subjected to a publicity constraint itself is to demand that we endorse a principle prior to selecting it.
destination, I shall then consider whether some factors should be weighed more heavily than others in determining which option to pursue. It will also be necessary to explain how we might go about making trade-offs between the different sorts of values identified by the normatively significant features, given that it seems highly unlikely that we will ever be lucky enough to be deciding between just two alternative proposals that differ only with respect to one of their normatively significant features.

**Feasibility**

Consider two political proposals, equal in every respect except that one is much more feasible than the other (i.e. much more likely to be successfully implemented). It seems clear to say that, all else being equal, the more feasible the proposal, the more desirable it is. We can explain an obligation to choose the more feasible option by referring to an obligation similar to the one assumed by the due diligence requirement in Section 5.1. We must not expose people to unnecessary risks and to choose to pursue a less feasible option over a more feasible option necessarily violates this duty in any case where failure to implement a proposal entails that some people will be harmed (almost every case will entail this). We might think that the strength of the general obligation is sensitive to the costs of minimising risk, and to the harms caused by failure, but in cases where all else is equal, and the goal is normatively valuable, the higher the feasibility score, the better the proposal.

**Stability**

There are two ways we might understand the concept of stability as it applies to political proposals. First, we might describe a proposal as more stable than another to the extent that it is more consistent and reliable. Suppose, for example, that our goal is to minimize infections of a particular disease and that there are two policies available to us, each of which requires investing in and distributing a particular drug. Suppose that if we choose Drug A, it will tend to cure people 90% of the time, but Drug B will tend to cure people 99% of the time. We may regard Drug B as offering the more stable option to the extent that it offers a more reliable solution to our problem. Suppose also, however, that people who take Drug B have a 50% chance of developing a resistance to Drug B, such that after a period of five years the drug loses its effectiveness. In contrast, suppose that Drug A, while only curing 90% of people, tends to cure them permanently. In this sense, Drug A offers a more stable solution, because it offers a more durable solution. In what follows, I will be mainly considering stability in this second, durability-implying sense.
Whether or not stability should be thought of as a relevant factor in deciding between political proposals depends upon how the destination is defined. Suppose, for example, we have one political proposal that would have the effect of eradicating world hunger for one week, and a second political proposal that would have the effect of eradicating world hunger for the foreseeable future. If we define the ideal as ‘eradicating world hunger’, we might view the first proposal as an unstable one which would achieve the ideal, but only temporarily.

If that is the right way to conceptualise the proposals, then stability is one of the normatively significant features of a proposal that matters when we are selecting between different routes to the same destination.

There is an alternative way to conceptualise the two proposals, however, such that we might think that each is aiming at a different destination, rather than one destination being merely an unstable variant of the other. On this view, the ideal might be permanent eradication of world hunger, with a second (or third, fourth etc.) best solution being the eradication of world hunger for one week. On this view, stability plays a role in destination-selection rather than route-selection.

The value of stability will likely change depending on whether we are considering it as a property of a route or a destination. If we determine that some possible destination is desirable but unstable, for example, this might give us decisive reasons to change the ideal that we’re aiming at. On the other hand, if we are considering the stability of a possible route to a stable destination we may be willing to gamble on an unstable route that has other attractive features (such as involving little additional risk in case of failure, for instance). So, when looking at destinations, stability may be a necessary condition for regarding it as (evaluatively) desirable, while stability may be less important when considering the normative value of a route given that the disvalue of an unstable route might sometimes be outweighed by other normatively significant properties that the route in question might possess.

**Side-effects and efficiency**

When comparing two political proposals that aim at the same (stable) destination we ought to prefer proposals that are most efficient, all else being equal and proposals that produce the most desirable side-effects, all else being equal. Less efficient proposals waste resources,
which could be used to make a second attempt, should the initial attempt to implement a proposal fail, or which could be used to implement political proposals aimed at alternative destinations. Proposals that produce beneficial side-effects at no additional costs to implementation are to be preferred over those which produce none, or which produce fewer beneficial side-effects, for reasons I take to be obvious.

‘Deep’ moral constraints

When it comes to principles that determine destination selection, I suggested above that we may be required to deviate from the pursuit of the ideal in cases where this is necessary to comply with some overriding ‘deep’ moral constraint. Similar constraints apply at the route-selection level, such that political proposals that would have the effect of achieving the desired goal while violating underlying foundational moral constraints must be rejected in favour of the set of proposals that offer morally permissible paths to the objective. These principles can articulate basic deontological or consequentialist principles, for example, or they may express more complex ideas, such as the notion that proposals that can satisfy a publicity criterion are to be preferred over proposals that cannot.

This category of deep moral constraints is extremely broad (and vague), because it covers any moral considerations that come from “outside” the context of the particular routes and destinations in question. The idea here is that once we have assessed and selected a particular proposal aimed at achieving a desirable ideal, we must then check it against our wider views about moral principles in order to ensure that the proposal is still worth pursuing, all-things-considered

Path Dependency

Certain destinations, or states of affairs, are only accessible via certain routes, and certain political proposals require closing off potential future routes if we attempt to pursue them. We can thus create hard constraints that make particular destinations permanently inaccessible by moving to a status quo from which no available path to the ideal remains.

Where this kind of path dependency is an issue, or where there is only one available path to a particular destination, it is vital that our evaluations include not just the status quo and potential future states of affairs, but the entire sequences of events between the two.
Where multiple paths are open, we should prefer political proposals that maintain a degree of openness, allowing us to try alternative approaches, or to make repeated attempts to achieve a destination, when all else is equal.

5.4 Tensions and Trade-offs

In the previous sections, I offered an account of the normatively significant features of political proposals that should be decisive when it comes to choosing between different routes or different destinations, all else being equal. In practice, it is difficult to imagine any case where we are faced with a choice of routes or destinations that differ only in respect to one of these features. Having given an account of the features of a proposal that are normatively significant, I now turn to the much more difficult question of how we should weigh each particular feature of a proposal, such that we can make the right decisions when we are forced to make trade-offs between different normatively significant values. I shall begin by considering cases where we are faced with a range of political proposals aimed at achieving a particular ideal, i.e., where we are engaged in the route-selection process.

One way we might distinguish between the different features is to assign lexical priority to one or more such that proposals are first judged according to the lexically superior feature and only in cases where proposals are equal in this measure (or where each proposal passes a minimum threshold with regard to the feature in question) do we then look at additional normative features of the proposals under consideration. Are any of the features listed previously the sort that warrant this kind of status?

Feasibility in its binary sense (where it distinguishes possible proposals from impossible ones) might seem to be a good candidate for lexical priority, since we might think that a proposal that is straightforwardly impossible to implement should always be rejected regardless of its other features. As discussed in Section 1, however, the pursuit of an ideal that is impossible to achieve can also be valuable (perhaps because it produces valuable consequences, or because its pursuit is non-instrumentally valuable), and it is possible to imagine a case where the best option before us is to pursue an infeasible (in the binary sense) political proposal. Feasibility in its non-binary sense appears to fare no better; we can imagine cases where a more feasible but less stable proposal is rejected in favour of a more stable (but less feasible) proposal, but we could also imagine a case where a more stable but less feasible proposal is rejected in favour of one that is more feasible but less stable. If there
are features of political proposals that are of paramount importance and are antecedent to all additional considerations, feasibility (in either its binary or non-binary sense) does not look to be a viable candidate.

In order assess the role that feasibility should play when it comes to trade-offs, we need to begin by looking at the sorts of reasons we might have for preferring more feasible proposals in certain cases, and the sorts of reasons we might have for rejecting more feasible proposals in favour of less feasible ones in other cases. When we opt for more feasible proposals (when considering alternative routes to the same destination) this is likely to be in response to prudential reasons to do with using resources efficiently and maximising the chances of success as well as normative reasons to do with obligations to avoid exposing people to unnecessary risks and harms. If other considerations can trump feasibility, such that we may be permitted or required to pursue a proposal that is not the most likely to succeed, what might those values be?

Stability as a value appears to be clearly contingent on other values, given that whether or not stability is valuable at all depends upon the destination in question. Stability is perhaps best thought of as serving an amplifying function such that it makes desirable destinations more desirable and undesirable destinations even less desirable. Similarly, lack of stability makes desirable destinations less desirable and may make undesirable destinations more desirable.\textsuperscript{72} Not only does stability appear to be empty of normative significance except when it applies to an ideal which is normatively significant for other reasons, prioritising it as a foundational value would be risky in the sense that it would make us much more liable to defer to a stable status quo over a less stable ideal. We might consider combining stability with another value, but it would still be possible to imagine cases where such a combined value could be trumped by another. For example, a highly feasible and stable proposal might nevertheless be less desirable because of path dependency issues, or a proposal that would be highly stable and free of path dependency issues might be less desirable because it has an extremely small chance of success.

\textsuperscript{72} I am less certain about the extent to which we should say that a lack of stability will necessarily make an undesirable situation more desirable. We would presumably need to say a little more here about what lack of stability means in the context in question. If, for example, it indicates that the undesirable state of affairs will only last a short time, or that it will be easier to escape than a more stable state of affairs, then it seems plausible to say that we should view the unstable and undesirable destination as being better than its stable equivalent, all else being equal.
Perhaps we might want to prioritize those options that, if chosen, would be likely to keep other options available, should our chosen path turn out to be inadequate. Let’s call this the value of ‘path-dependency’ in the sense that a proposal may be more valuable to the extent that it does not restrict our abilities to choose a different path or route, in the future. Path-dependency as a value looks to be primarily sensitive to issues of feasibility, given that we have less reason to be concerned about path-dependency in cases where we are confident that our proposals will succeed and that the destination we are aiming at is worth pursuing. Prioritising path-dependency might seem to be the best way to satisfy the requirement that we not expose people to unnecessary risks, since we would always be keeping our options open to the greatest extent. However, the obligation to minimise risk is itself not the only normatively significant obligation we must consider—sometimes we must meet deeper moral obligations by taking risks, for example. Even if there were no other obligations except to minimise risk, it is not clear that prioritising path-dependency as a value would necessarily minimise risk in every case. Sometimes being overly cautious and conservative can be a riskier strategy in the long run.73

The only remaining candidates for lexical priority are what I have called deep moral constraints. It is important here to avoid begging the question by simply defining these as the sorts of things that must be prioritized at the expense of each of the other values. I shall consider three kinds of deep moral constraints; consequentialist constraints, deontological constraints, and constraints imposed by a political liberal conception of justice.

If we take our deep moral constraints to be rule or act consequentialist, then we must regard each of the values discussed previously as having instrumental value such that, when forced to make trade-offs between them we are guided by a core consequentialist principle that determines which political proposal we choose. An act consequentialist view might suggest, for example, that we always choose the proposal that is likely to have the best consequences overall, where this is defined according to whatever metric of the good our consequentialist theory endorses. A slightly more complicated rule consequentialist approach might say instead that we ought to adopt a procedure for making trade-offs that is likely to have the best results overall, even if this will sometimes mean choosing a sub-optimal (in the sense of the direct consequences it produces) political proposal.

---

As I noted earlier, it may sometimes be the case that closing off options is a useful way of minimizing risk.
What I have in mind here is that one consequentialist view might regard each proposal selection in isolation, such that only the potential consequences of the proposal in question count when evaluating its desirability. A different kind of consequentialist view might instead emphasise the decision procedure that ought to be used for selecting political proposals more generally, such that individual proposals may sometimes be selected which do not produce the best consequences, provided that the rules for selecting proposals in general tend to lead to the best consequences. So, for example, a view of the former kind may not constrain a proposal that seeks to achieve an ideal by sacrificing the lives of a minority to improve the lives of a majority, while a view of the second kind may constrain such a proposal on the basis of standard rule utilitarian arguments against act utilitarianism.

If the deep moral constraints have a deontological rather than a consequentialist flavour, then we could imagine them as serving a kind of ‘ruling out’ function, such that, when faced with a potential trade-off between values we must begin by excluding any options that require us to violate the deontological maxims that form the core of our underlying moral theory. If we are constrained by a maxim that tells us never to violate certain strong principles of self-ownership, for example, then any conflict between self-ownership rights and some other value, such as feasibility, stability or path-dependency, must always be decided in favour of the values protected by the deep moral constraints – in this case, the value of self-ownership – even if this would lead to worse consequences overall.

Given that the focus of this chapter, and indeed the thesis more generally, is on political proposals, we might think that the deep moral constraints should be especially sensitive to the fact that the proposals will be implemented in the political sphere. In cases where trade-offs must be made, this view says that we must never choose options that could not be accepted by all reasonable persons who shall be affected by the implementation of the proposal in question. What should count as a ‘reasonable person’ is somewhat contentious, and I shall not attempt to address that difficult question here – it should be sufficient to assume that there can be some plausible account of what makes a person reasonable. This approach looks as though it will yield an account of deep moral constraints that is more restrictive than the basic consequentialist and deontological approaches outlined above. For example, there may be cases where it is clear to some people what option a consequentialist or deontologist ought to choose, but it is not possible to communicate all of the necessary facts to the group of reasonable people who will be affected by the proposal (or that even if such facts are communicated, the issues will still be the subject of reasonable disagreement). If there are
alternatives, which are less feasible, less stable, more path-dependent etc., but which are nevertheless able to withstand the requirement that they be subject to acceptance from all reasonable people, then those alternatives are to be preferred, even if the publically-inaccessible proposal would yield better consequences overall, or would do a better job of adhering to deontological maxims.

Even if deep moral constraints deserve lexical priority, be they based on one of the three approaches sketched above, or some alternative, the more general problem remains largely unresolved, for two reasons. Firstly, while deep moral constraints will rule out some proposals, or decide some cases where values conflict, there will remain many cases where deep moral constraints are unable to decide between proposals which involve trade-offs in other values (stability, feasibility, etc.) but which do not threaten the values affirmed by the deep moral constraints themselves. For example, suppose that we have a range of publicly accessible options which would each adhere to some fundamental set of deontological maxims, but which would nevertheless require us to choose to implement either a proposal that is more stable, or a proposal that is less path-dependent. In such cases, we begin choosing which proposal to implement by taking the set of all available options and ruling out those that would violate deep moral constraints, but we remain faced with a subset of options that satisfy the deep moral constraints yet still require that additional trade-offs be made. In such cases, assigning lexical priority to deep moral constraints will be a necessary but not sufficient part of choosing the correct political proposal.

There are other ways to make trade-offs that do not require us to assign a fixed lexical priority to values, however, and we might decide to opt for one of these methods once we reach the point where, having ruled out as many options as we can, given deep moral constraints, we are nevertheless faced with difficult trade-offs between competing proposals.

For example, we could try an aggregative approach, such that, when faced with two political proposals that have each passed the deep moral constraints test, we simply compare each proposal's normatively significant factors to its rival and choose the proposal that "wins" the most number of comparisons. So, for example, Proposal A might be more stable than Proposal B, as well as being more feasible and less risky, but Proposal B might be less path-
dependent. Here we might say that Proposal A defeats B three points to one. The problems with such a crude approach should be immediately apparent.\footnote{In case they are not: there will be cases where the score between proposals is tied, and it seems implausible to think that in such cases we ought to be indifferent as to which proposal we should choose. While any plausible account of trade-offs must allow for cases where we should be indifferent between proposals, these cases are not likely candidates. For example, it is plausible to think we should be indifferent between a proposal which will save five innocent people in location A while allowing five people in location B to die, and a proposal which would save the five innocent people in location B, while allowing the five innocent people in location A to die, all else being equal. But it’s not at all clear that we should be indifferent to a proposal which is more feasible and stable and another proposal which is less path-dependent and less risky.}

If any kind of weighted aggregative approach might work, such an approach would have to be context-sensitive while also being consistent enough that we would be able to articulate and follow some set of principles that tell us how we should go about assigning weight to different values in general. It might seem as though some kind of consequentialist principle might be up to this task – the principle might say something like “having taken deep moral constraints into account, when considering the proposals that remain, we ought to assign the sorts of weights to different values such that we will tend to produce the best outcomes by doing so”.

This is not likely to be a successful approach however, for the following reason. In order to know what outcomes are most desirable, we need to be able to assign a certain weight to the value generated by each of the normatively significant features of the proposal. We need to know whether it tends to be better to weigh feasibility more highly than stability, or to weigh path-dependency more highly than riskiness and so on. This, however, presupposes precisely the sorts of capacities we're trying to use the principle to generate, namely, the capacity to assign a consistent and non-arbitrary weight to the normatively significant features of a political proposal. Nor does it look as though it will be possible to reduce the goods generated by the different normatively significant features into a single value such that the suggested consequentialist principle will be able to give us the sort of guidance we need.

It is possible that sometimes there will simply be no (evidence-relative) fact of the matter as to which proposal we ought to choose, given the evidence available to us. In such cases, while we may well have reasons to regret being placed in such a situation, and while we may or may not be culpable for being in such a situation, how we ought to act may seem to be a

\footnote{Secondly, and more importantly, even in cases where the 'score' is not tied, there is no good reason to think that each value should always count as much as any other value. As I suggested above, sometimes feasibility will be more important than stability, sometimes path-dependency will be more important than riskiness and so on. Applying a fixed weight to each of these values would bring us back to where we began in trying to assign lexical priority to each one.}
relatively straightforward matter. Suppose, for example, we have good reasons to believe that Proposal A is equal to Proposal B in every respect except that one of the proposals is less feasible, but we don't know which one of the two is less feasible. Or suppose we adopted a principle which tells us that we always ought to prefer a proposal that is more feasible and stable over a proposal that is less risky and less path-dependent, but we don't know whether Proposal B is less risky than Proposal A. Or suppose we know that Proposal A is more stable and feasible and less risky and less path-dependent than Proposal B, but we don't know whether Proposal A violates deep moral constraints.

In each of these examples there may be facts we could uncover that would give us the right answer (given the particular principles we have adopted), but if we have good reasons to believe that those facts are inaccessible to us and if we must make a decision about which proposal to choose, we might think that we could simply flip a coin, or take a majority vote, and implement whichever proposal is chosen via this procedure. This would sometimes mean that we choose the suboptimal proposal, but, so the thought goes, it's the best we can do, given the limited information available: we should expect mistakes to be made when we make decisions without all the information.

There is an additional complication here, however, such that it may turn out that when we make decisions without all of the information we have reasons for tending to prefer certain kinds of proposals over others. We might think, for example, that deep moral constraints are so important that we ought never risk implementing a proposal unless we are reasonably certain that it is consistent with deep moral constraints. Or we might think that we ought to prefer proposals where we are certain about how risky they are even in the face of proposals that are potentially more desirable on other measures, but where we are uncertain as to the risk. These sorts of questions force us to make very similar decisions about trade-offs and values that we were making in the cases where all of the relevant information about the proposals' competing features were available to us. I see no more reason to be optimistic about our ability to articulate principles about making these kinds of epistemically constrained decisions than we have to be about our ability to make decisions between proposals where epistemic constraints don't apply.

If these are intractable problems, or if these are problems we don't believe will be solved any time soon, it remains for us to decide what we ought to do in the face of such problems. Given epistemic constraints and the lack of a solid conceptual framework upon which to
make decisions between certain kinds of proposals (i.e. proposals that are consistent with deep moral constraints, but which vary across their different normatively significant attributes such that we have no clear intuitions about how we ought to act) we require a non-ideal approach that can tell us how we ought to act in the face of such constraints. What would such a non-ideal approach be likely to recommend?

The correct non-ideal approach needs to possess certain features in order to give us consistent, implementable (i.e. action-guiding) and comprehensive recommendations across a large range of cases. First, we need to know how we ought to act in cases where we are faced with a potential future uncertainty (as to how to make trade-offs), but where we have good reasons to believe that there are actions we could take which would reduce or eliminate that uncertainty and place us in a position either to choose one particular proposal, or at least to narrow down the range of options under consideration. As these cases deal with soft constraints, let's refer to them as ‘soft cases’.

The best non-ideal approach also needs to be able to deal with ‘hard cases’, i.e. cases where it is foreseeable that we will be unable to choose between a particular range of options, and that there is nothing we can do to avoid this problem.

In addition to these, there will be two kinds of what we can call ‘indeterminate cases’. The first kind of indeterminate cases are those where we are unable to determine whether the constraints we are dealing with are hard or soft. The second kind will be cases where we are able to determine whether the constraints in question are hard or soft, but we are unable to determine the particular strengths of important relevant constraints.

In order for the correct non-ideal approach to be a comprehensive one, it needs to have enough to say about what we ought to do in each of these cases. In order for it to produce recommendations which can guide our actions, it needs to begin by offering some clear principles that explain under what circumstances we ought to act, even at the risk of doing more harm than good, given uncertainty about how we ought to make trade-offs between competing political proposals.

Recall that in Chapter 4 I suggested that we ought to respond to the constraints that apply to theorists as well as their theories by engaging in a ‘non-ideal’ form of ideal theory. In that section, I used the metaphor of a chess game to describe how we might go about developing decision-making capacities to reduce the number of available ‘moves’ in order to make it
more likely that we succeed in our overall aim of choosing the best option available. I believe this sort of approach can serve as a blueprint for a non-ideal approach to transitional theory as well.

The first step in constructing this position is to make an important concession which will be central to the approach and which defines it as a thoroughly non-ideal response to the problem in question. This concession is the starting assumption that we are and will be, at least for the foreseeable future, unable to articulate and defend principles that will offer us clear guidance such that we will be able to choose the best proposal with any degree of confidence when dealing with hard cases.

Setting such cases aside for the moment, we have reason to be more optimistic about our ability to deal with soft cases, provided that we focus on our ability to develop our capacities to identify, assess the strength of, and overcome soft constraints. If soft constraints are not dealt with in time, or effectively, they can become hard constraints once we lack sufficient resources to overcome them.75

Consider the case of global climate change, for example. It is a generally agreed-upon fact that we possess sufficient resources to prevent global climate change from doing irreparable harm to the environment, provided that we act effectively and in time. If the issue is not addressed effectively and in time, there will come a point at which we no longer have sufficient physical resources to deal with the problem. As this point approaches, we will be faced with difficult trade-offs that will grow increasingly difficult to make as hard cases proliferate.

When it comes to transitional theory rather than ideal theory, what matters in particular are the conceptual and epistemic resources, rather than other kinds of physical resources, which can prevent surmountable problems from becoming intractable in the long run. Suppose, for example, that we believe that we will have to make trade-offs between the interests of future generations and present generations when it comes to deciding on a strategy to combat global climate change. Some of these trade-offs may be inevitable and unsolvable with any degree

75 It might be objected that it should be easier to deal with hard cases, rather than soft, since we will have fewer options that are feasible and thus less complex choices when dealing with the former. In contrast, when dealing with soft cases, there is more opportunity to make errors of judgment, given that we have to engage in a broader array of decisions about how to weigh competing variables, and so on. This may be true initially, prior to our deliberation about either hard or soft constraints, but what distinguishes hard cases from soft is that the former require us to deal with constraints which are in principle irresolvable. While the sheer number of questions which require answers when dealing with soft constraints does increase the difficulty of finding the right answers, at least those answers are accessible in principle.
of confidence. Others, however might be avoidable and/or solvable, provided that we develop certain capacities. For example, suppose that one reason we are uncertain about how to make trade-offs between current and future generations is that we are uncertain what the needs of future generations will be. It might be that certain kinds of research (in environmental development, healthcare, social science and so on) would enable us to make predictions about the needs of future generations with greater accuracy, which would help us to rule out at least some options when it comes to making trade-offs. Or suppose that one reason we end up having to make more trade-offs is because we are using our physical resources inefficiently such that a more efficient use of resources would avoid us having to make certain kinds of trade-offs in the future. Both strategies – seeking to avoid having to make trade-offs and seeking ways to make better-informed trade-offs are ways of minimising the problem of hard cases by dealing with soft cases before it's too late.

Given the above, I suggest the following principle we ought to endorse:

**The Principle of Optimization:** We ought to try, within reason, to minimise the chances that we will have to make trade-offs in hard cases by maximizing our capacities to identify, assess the strength of, and overcome the epistemic constraints which make it more likely that we will have to make trade-offs in hard cases.

The ‘within reason’ clause in the above principle stands in need of clarification. What I have in mind here is that there must be plausible limits as to the point at which we are entitled to stop investing further resources (whether mental or physical) in developing our capacities to identify, assess and overcome constraints in pursuit of the ideal.

One way to do this might be to try to work out the expected utility of continuing to develop the relevant capacities and weigh that against the expected utility of ceasing, or lessening the pursuit of the capacities in question. Of course, this in turn presupposes a capacity to make sufficiently accurate assessments of expected utility in both cases.

At the very least, we ought to pursue the development of the relevant capacities to the extent that we are actually able to identify ideals that are plausibly worth pursuing, and to the extent that pursuing the development of the relevant capacities does not significantly impact upon our ability to address the issues we have identified.

What this means is that we need not worry about trading off future development of the relevant capacities until we have actually identified something that this must be traded off
against, and that we obviously need not worry about those cases where developing our capacities can be pursued at negligible cost to the pursuit of an ideal.

Fully adhering to these very weak requirements may nevertheless constitute a significant improvement over the status quo.

5.5 Conclusion

One of the reasons it is important to understand the relationship between ideal and non-ideal theory is so that we can understand how we might begin to move from a non-ideal status quo, to a more desirable state of affairs and, ideally, towards the best possible state of affairs.

One of the reasons that the Constraint-Based View is useful is that it provides us with a fairly straightforward way of imagining this process: moving from a non-ideal status quo requires offering an accurate account of the relevant constraints that apply to the status quo, an accurate account of what a world would look like if only the relevant hard constraints applied, and coming up with proposals aimed at removing the soft constraints that stand between us and a more desirable state of affairs.

One way to evaluate these proposals is to ask whether they constitute a deviation from the pursuit of the ideal and, if so, whether that deviation can be justified in the case in question. Transitional principles are those principles that tell us whether deviation from the pursuit of an ideal is justified. In this chapter, my aim has been to sketch what these principles might look like.

First, I identified the following principles that apply when considering questions of destination selection:

**TP* (Impossibility):** If we have sufficient reason to believe that a particular ideal is not achievable, both now and at any point in the future, and if ought implies can, and if the pursuit of the ideal is not more valuable than any alternative proposal under consideration, we cannot be obliged to pursue that ideal and it is therefore at least permissible to deviate from the pursuit of that ideal.
TP2 (Due Diligence): In order to deviate from or to pursue an ideal, all reasonable efforts must be made to ensure that the decision is justified. This entails taking measures to ensure that we have sufficient capacities to ascertain whether a proposal is achievable, to what extent it is desirable and that we have sufficient capacities to apply transitional principles themselves.

TP3 (Demandingness): It is permissible (but not required) to deviate from the pursuit of an ideal if it is the case that the pursuit of the ideal can be reasonably expected to impose costs upon agents which are overly demanding, such that pursuit of the ideal would ask agents to bear costs which are greater than we believe the correct theory of morality requires.

TP4 (Deep Moral Constraints): Deviation from the pursuit of the ideal is permitted and may be required if the pursuit of the ideal would require the violation of some deep moral constraint, where this is understood as a fundamental normative commitment, the content of which is determined by the broader moral theory we believe to be correct.

TP5 (Publicity Constraint): Deviation from the pursuit of the ideal is permissible provided that the reasons for the deviation are ones to which all reasonable people could assent. These reasons may include the fact that pursuit of the ideal itself cannot be defended upon grounds to which no reasonable person could reject.

Following this, I focused on some of the most normatively significant aspects of routes that must be addressed when answering questions of route selection. These aspects included questions of feasibility, stability, side-effects and efficiency, path dependency and deep moral constraints. I focused in particular on cases where it is difficult to know how to make trade-offs between these different values, and suggested the following principle as a response to these difficulties:

The Principle of Optimization: We ought to try, within reason, to minimise the chances that we will have to make trade-offs in hard cases by maximizing our capacities to identify, assess the strength of, and overcome the epistemic constraints which make it more likely that we will have to make trade-offs in hard cases.
This principle might be seen as a kind of non-ideal response to epistemic constraint surrounding route and destination selection, and when combined with the five transitional principles outlined above, should help us to get a sense of what a good approach to transitional theory would look like, though I have only been able to offer a sketch of it in this chapter.
6. The Constraint-Based View in action: Justice for Future Generations

6.1 Introduction

If the Constraint-Based View of ideal and non-ideal theory (hereafter, ‘the CBV’) is to be useful as a conceptual model, there should be clear advantages to using it when considering at least some contemporary debates about justice in normative political theory. Having described and defended the CBV in the previous chapters of this thesis, the aim of this chapter is to offer an example of how the CBV might be applied to a particular set of questions about justice, in order to illustrate its usefulness as a conceptual model. The case I have chosen to consider in this chapter is that of intergenerational justice, though there are presumably numerous other cases to which the CBV could apply. I have chosen to focus on intergenerational justice in particular because it involves many different kinds of constraints that are related to one another in relatively complex ways. If the CBV can offer a coherent and useful way of assessing these issues, then I believe we should be optimistic about applying the view to other issues of justice.

In Section 6.2, I shall offer a brief, descriptive account of three issues that are central to questions of intergenerational justice. My aim in this section, and throughout this chapter, is not to argue for answers to any of these questions (that would be far beyond the scope of this thesis). Rather than providing the answers directly, the CBV is supposed to make it easier for us to answer the relevant questions, which first requires an explanation as to what these questions actually are: this is the purpose of the first section.

In Section 6.3, I shall apply the CBV to the issues considered in Section 6.2. This section shall also be largely descriptive, concerned primarily with translating questions of intergenerational justice into the language of the CBV.

In Section 6.4, I shall argue that the analysis that emerges following the application of the CBV offers three distinct benefits. First, I will argue that the CBV gives us clear and helpful prescriptions about how we should conceptualise the central questions of intergenerational
justice, such that we are able to assess the issues with a greater degree of clarity than before. Second, the CBV, via the ‘narrowing down approach’ discussed in Chapter 4, can offer us a clear path towards a non-ideal theory of intergenerational justice. Finally, in this Section, I will argue that the CBV, via the transitional principles outlined in Chapter 5, can offer a path towards the ideal, by specifying the conditions that need to be met in order for us to deviate from pursuing a world where our obligations towards future generations are fully realised.

In Section 6.5 of this chapter, I offer a brief sketch of what a ‘non-ideal’ approach to intergenerational justice could look like, using the CBV as a conceptual framework.

6.2 Describing the Debate

Most people believe that we have moral obligations towards people who will be born after we die, given that some of the decisions we make in the present can have profound effects upon future generations not only in terms of whether the lives of future people go better or worse, but in terms of which people and how many people will exist.

Specifying the structure, scope, and content of these obligations is obviously an important task – we need to be able to tell whether we are meeting our obligations to both future generations as well as the present generation, which will have to make do with fewer resources or opportunities if future generations are to be provided with more. However, figuring out what we owe to future generations has proved to be especially tricky, given that the philosophers who have been considering these issues for the last forty years or so have been unable to reach a consensus on whether we have any moral obligations to future generations at all, let alone what those obligations might consist of.

In this section, I will focus on four related problems that arise once we consider these issues: (1) The ‘non-identity problem’, (2) The ‘repugnant conclusion’, (3) The feasibility of predicting the consequences of our choices in the present for people in the future (4) The possibility (and desirability) of being motivated to act to benefit the lives of future generations. In Section 6.2, I will show how these problems can be described using the language of the CBV.
The Non-identity problem

When considering the scope and structure of moral obligations to future generations, the chief difficulty has not been disagreement over the intuition that we do have moral obligations to future generations – most people, philosophers and non-philosophers alike, share that intuition - but rather in articulating a plausible principle or principles that can ground these intuitions.

The identification of the ‘non-identity’ problem marks an important turning point in how philosophers have tended to think about obligations to future generations. Initially, many philosophers focused on the fact that future people are not yet alive, and framed the issue in terms of whether we can act wrongly towards people who have yet to be born. However, this way of thinking about future people treats them as though the existence and quality of life of future people is in a sense inevitable, as opposed to being wholly contingent upon the acts of those of us alive in the present.

Gregory S. Kavka's “The Paradox of Future Individuals” and Derek Parfit's “Future Generations: Further Problems” were particularly important in challenging these assumptions. Kavka points out that it is not merely the interests of future persons that the present generation can affect, but the very existence of those persons themselves. Provided that one accepts some relatively uncontroversial assumptions about personal identity it is possible for us to create future persons who live lives barely worth living, and whose suffering could not have been otherwise (i.e. we could not have created those same people who live lives of less suffering). Here is one of the cases Kavka uses to illustrate this point:

Imagine a society choosing between heavy investment in either solar energy systems or nuclear fission power plants in order to meet projected energy needs over the next few centuries. Monetary investment costs per unit of output will be slightly lower if the nuclear option is pursued. However, there is no storage or disposal system for nuclear wastes that is expected to contain them safely for more than few generations. Hence, if the nuclear plan is

---

76 Chronologically speaking, the trouble began around the 1960's with questions about how utilitarian theories of morality should regard obligations to future generations; especially with regard to whether utilitarianism requires us to have as many and as happy children as possible, in order to meet our utilitarian duties to maximise happiness. So, for example, Jan Narveson, in his 1967 paper “Utilitarianism and New Generations” is mainly concerned with demonstrating that utilitarianism neither imposes a moral obligation on us to have children (with a few minor exceptions), nor permits us to produce children whose lives would not be worth living. In “The Rights of Animals and Unborn Generations”, published in 1974, Joel Feinberg considers whether future generations can be the bearers of rights against members of the present generation, who have it in their power to affect the interests of future generations in important ways.

77 Specifically, one needs to accept something called the “genetic identity view”, which says, the identity of a person is at least in part constituted by the DNA the person has as a result of which ovum was fertilized by this or that spermatozoon in the creation of this person. See for example Parfit, (1984) pp. 355-356.
pursued, there would very likely be radioactive leakage that would cause the deaths of thousands of people in future generations. Intuitively, it is clear that it would be wrong to build the nuclear plants rather than the solar energy systems, as doing the former would lead to the premature demise of thousands. But, because the choice of the nuclear plan could affect conditions for conception nationwide (relative to the selection of the solar plan), these thousands of individuals would never have existed had the solar energy option been pursued instead (Kavka, 1999, p.97).

Examples like this one, developed further by Parfit in *Reasons and Persons* in 1984, form the basis of what became known as ‘The Non-Identity Problem’. On the one hand, almost everyone has the very strong intuition that it would be morally wrong for a society to pursue the nuclear plan, described above, yet the fact that the people who would be harmed by the nuclear plan could only exist in the first place if the nuclear plan is produced, seems to render the nuclear plan permissible. This, in turn, seems to be an absurd conclusion, given our initial intuitions.

Engaging with the non-identity problem requires engaging with a cluster of related questions, many of which are extremely difficult to answer. Some have focused on the plausibility of Parfit's account of benefiting and harming, upon which the non-identity problem rests, but these are not the only questions that need to be answered if the problem is to be fully solved, and if our obligations to future generations are to be made clear.

Aside from more metaphysical questions of personal identity, and whether bringing a person into existence can benefit them, or whether we can harm people who would have been ‘better off’ had we not acted, central to the discussion of the non-identity problem is the notion of a life that is ‘worth living’. The non-identity problem presents itself most sharply in cases where there is the possibility of creating people whose lives, while deficient in some respect, are on the whole still worth living. Articulating a plausible set of principles to determine when a life counts as worth living in this sense is obviously another large question that would need to be answered as part of a comprehensive account of our obligations to future generations. So, even solving the core question posed by the non-identity problem (how can we have obligations to people whose existence is contingent upon the very actions

---

78 Specifically, what seems to render the nuclear plan permissible is that there is no counterfactual situation where it is true both that the future people exist and that they are better off than they would be under the nuclear plan. Thus on a counterfactual understanding of harm, it seems as though we fail to harm them.

79 For example, James Woodward (1986) argues that in cases such as the nuclear plan, the original planners are not entitled to ‘take credit’ for the lives of the future people, even though their actions played a necessary causal role in allowing them to lead lives worth living. Similarly, Matthew Hanser argues that future people who suffer harm as a result of things like the nuclear plan are harmed in a morally relevant sense, and that the original planners don’t deserve credit for benefiting them, merely by bringing people into existence in a way they almost certainly could not foresee.
that we supposed to be morally problematic?) will not in itself be sufficient to give us a complete account of what we do and do not owe to those who have yet to be born.

Some philosophers have been tempted to respond to the non-identity problem by focusing on the fact that the problem seems to arise when we make our moral principles sensitive to the existence and interests of particular individuals whose existence and quality of life in the future is determined by our actions in the present. However, if we make our moral principles impersonal, such that they track the existence of value in general, it looks as though the non-identity problem can be avoided.

To see why an impersonal approach might help us avoid the problem, suppose that we are considering whether we may act in a way that will create future people whose lives are barely worth living, or whether we ought to act in a way that creates the same number of different people with better lives. If we take the view that our decision should be sensitive to the harms and benefits we would cause to particular individuals, this would normally be the point at which the non-identity problem arises. Whomever we choose to create will seemingly have no cause for complaint against us, provided that their lives are at least worth living.\(^{80}\)

However, suppose we adopt a view that says that what matters is not whether particular people are harmed or benefited by our actions, but rather what matters is how much value a world contains, considered from an impersonal perspective (this is a very crude utilitarian view, but it should suffice for the moment).

On this view, we have a clear obligation to create the people with better lives, rather than the people whose lives are barely worth living, because this will lead to a world that contains more value (bearing in mind that we are dealing with the same number of people in either case). Crucially, because we are considering value from an impersonal perspective it does not matter who actually experiences this value, so we avoid the non-identity problem altogether.

---

\(^{80}\) In other words, if a future person complains that their lives are barely worth living, we can simply reply that if we had acted otherwise, they would not have been born at all (someone else would have been, or nobody would have been).
The Repugnant Conclusion

While the above is an example of what Parfit calls 'same number' cases (Parfit, 1984, p.360), where we must choose between options which will bring about the same number of people whichever option we take, our obligations to future generations will also require us to engage with 'different number' cases, where our choices will not only determine which people come into existence, but how many people come into existence. At this point, adopting the impersonal approach leads to an extremely difficult problem in the form of ‘The Repugnant Conclusion’. As originally formulated by Parfit the Repugnant Conclusion is as follows:

For any possible population of at least ten billion people, all with a very high quality of life, there must be some much larger imaginable population whose existence, if other things are equal, would be better even though its members have lives that are barely worth living (Parfit, 1984, p.388).

It should be apparent why Parfit thinks we should find this conclusion 'repugnant'. Just as the non-identity problem seems to imply the highly-counterintuitive result that we might not have any moral obligations to future generations, provided that they lead lives which are minimally worth living, the Repugnant Conclusion seems to imply that we might be obliged to create vast numbers of people whose lives are barely worth living, provided that this would result in a world with higher levels of happiness or welfare (or whatever one's preferred measure is) from an impersonal perspective. If an appeal to impersonal principles is necessary to avoid the non-identity problem, then both of these problems combine such that the closer we come to solving one, the farther away we get from solving the other.

As with the non-identity problem, it is not necessary for the purposes of this chapter to delve too deeply into Parfit's arguments about the Repugnant Conclusion, or the attempts that have been made by Parfit and others to avoid the Repugnant Conclusion or to render it less intuitively unappealing. It is sufficient to point out that this remains a highly contentious area of debate, with no clear consensus as to how (or even whether) the Repugnant Conclusion is to be avoided.

Empirical Issues

Clearly, the substantive theoretical work with regard to obligations to future generations does not end once we figure out how to solve the non-identity problem while avoiding the repugnant conclusion. Such questions deal merely with the possibility of such obligations, but
say little if anything about what the content of those obligations should be. Whichever answer we settle upon, we will need, for example, an account of what makes a life worth living, as mentioned above. We will also need a comprehensive account of the distributive principles that shape our obligations – presumably these principles will come from our broader theory of distributive justice, and may perhaps be further modified according to the unique features of the relationship between presently living people and those who have yet to be born. Again, it is not necessary for the purposes of this chapter to further explore these issues, but it is important to note the sheer amount of work that must be done in order to provide a comprehensive account of our obligations to future generations. This will be especially relevant in the final section of this chapter, where I discuss how the CBV suggests we should act in the face of deep disagreement about such complex issues.

For the moment, let us set aside all of these difficult conceptual and normative questions about our moral obligations to future generations (including the non-identity problem and the repugnant conclusion). In figuring out what we owe to future generations, there will also be non-philosophical questions that will nevertheless be vital in producing recommendations about which distributive policies we ought to adopt. We can think of these as encompassing a broad category of empirical questions about the likely needs of future generations. The better our capacity to predict the needs of future generations, the less likely we will be to fail in our obligations to them.

These predictive capacities will ultimately need to cover virtually every aspect of the world that might have a bearing on distributive justice. Some obvious examples would be our capacities to predict when finite natural resources will be depleted, our capacities to predict the effects of future climate change, our capacities to predict (and preferably avoid) economic recessions and so on. As well as capacities to predict such events, we can also develop capacities to deal with unpredictable events, such as the emergence of new diseases or natural disasters, for instance. A comprehensive attempt to provide for the needs of future generations should include principles to deal with these sorts of ‘known unknowns’. That is to say, given that we can know with a certain degree of confidence that our capacity to predict future needs is limited in certain ways, we should build the knowledge of these limitations into the broader account of what we ought to do to ensure that the needs of future generations are met. Of course, unforeseeable developments may well include unexpected technological, economic or social developments which make much of our efforts towards future generations unnecessary or inefficient, and this is another factor which must be
incorporated into the broader account, but it would clearly be highly risky to assume that such developments are inevitable, or that they are inevitable before some arbitrary amount of time has passed.

Despite this need for caution, it does seem reasonable to suppose that we can assume that the farther into the future we go, the more difficult it will be to predict the needs of future generations. The only alternative to this view seems to be the assumption that at some point in the future, human development will eventually plateau such that the world remains in a relatively stable state in terms of people's social, technological and economic needs. This cannot be ruled out (indeed, it may be inevitable at some point) but predicting when and how such a plateau might be achieved seems to be at least as elusive as predicting how society might change in the distant future. In either case, it seems clear that our knowledge as to the needs of future generations will become less certain, the farther ahead we look.

This is going to be a problem for the general theory of obligations to future generations, especially if the correct theory tells us that the strength of our obligations to future generations does not correlate with their temporal proximity to us. Since our obligations will need to be sensitive to feasibility considerations, and feasibility considerations will in turn be sensitive to the confidence we have in our ability to predict the needs of future generations, it follows that our chances of meeting our obligations to future generations decrease as their temporal proximity to us increases.

While the distance in time between present and future generations looks likely to affect our ability to figure out what it is that future generations need and how best to provide it, temporal distance may create at least one major additional problem that threatens to undermine our ability to discharge our duties to future generations.

**Motivational Issues**

Roughly, the problem can be articulated as follows; in order for us to be sufficiently motivated to discharge obligations to future generations that require us to bear burdens here and now, it must be the case that we possess certain psychological dispositions towards future generations (including especially distant future generations). However, it is either (a) impossible or (b) undesirable that people should possess such dispositions, and this fact
presents a difficult if not insurmountable obstacle for any theory of obligations to future generations.

These ‘certain psychological dispositions’ are those attitudes that are necessary for us to feel compelled to act on our moral obligations towards future generations. To put it crudely, we might say that these dispositions concern a capacity to empathise with future generations to the extent that we feel compelled to meet our obligations to them in a way that is similar to how we feel compelled to meet our obligations to our contemporaries.

The idea that such a capacity is beyond the bounds of human nature seems patently false, given that there are at least some people who do seem to be perfectly able to motivate themselves to care about the interests of future generations. Perhaps one might think that this sort of attitude is exceptional, such that it is true in general of human beings (even though there are some exceptions) that we are not able to motivate ourselves to care enough about future generations to discharge our moral obligations towards them. I do not find this view convincing either – it seems to me to be much too quick and too pessimistic, given how malleable our motivational capacities seem to be in general throughout history.81

The idea that such capacities might be possible but undesirable seems to be based on the assumption that turning more of our attention to future generations might lead to a failure to discharge our duties towards present generations. While acknowledging that we bear additional obligations to future generations that we might not have previously admitted to will certainly complicate matters, I see no reason to assume that we will be unable to deal with these additional complications (and, of course, the fact that a view introduces new complications is not a good reason to reject the view if it is the right view). The only other reason I can think of why one might regard developing these capacities as being undesirable is if one believed that we don’t actually have the sorts of moral obligations that such capacities would suggest, in which case the real objection is to the underlying moral theory that suggests we ought to develop those capacities.

A related version of this problem has been discussed at length in the literature on global justice and moral cosmopolitanism, where the distance in question is geographical rather than

81 There are plenty of examples historical attitudes (sexism, racism, etc.) that have been firmly entrenched at one point or another but which have since been changed or substantially weakened. A bias against or indifference towards future generations (or the interests of non-human animals, or of geographically distant people) may well just be another form of motivational failure which most of humanity will eventually grow out of.
In that context, motivational objections tend to be feasibility-based and/or desirability-based. Feasibility-based objections challenge the idea that it is possible to achieve the kinds of global political institutions that would be necessary for people to discharge cosmopolitan obligations. There may be some overlap here between issues of global and intergenerational justice – perhaps there are certain needs (such as the need to live in a sufficiently clean environment) that would require international political institutions to deliver them for future generations. Desirability-based objections to cosmopolitan moral theories tend to focus on the idea that the sorts of moral obligations in question are incompatible with the partiality required by our bonds of co-citizenship. (Miller, 2009) Presumably, a similar objection might be made with regard to future generations (which might even involve distinguishing between future citizens and future non-citizens).

Given that global justice usually involves questions about obligations to actually existing people, it is in that respect less complicated than questions about future generations (though presumably figuring out what we do and do not owe to contemporaries will carry implications for bringing some people into existence and not others). Since questions about psychological dispositions persist in the context of debates about global justice, it seems plausible to think that they will present obstacles when considering temporal as well as geographical distance.

As noted above, the central problem can be divided into two questions – whether the necessary psychological dispositions are in fact possible, and whether they would be desirable if so. In figuring out the answers to these questions, it may be helpful to invoke arguments similar to those in the global justice debate. For example, when considering the possibility of cosmopolitan morality, Kleingeld and Brown (2013) point to the existence of various international organisations like the United Nations and European Union, and to the fact that some nations exist that contain vast numbers of people, often with a multitude of cultures, as evidence that it is possible to establish and maintain institutions that can be cosmopolitan in at least some limited sense. If we were to apply this strategy to the question of the possibility of meeting obligations to future generations, we could point to institutions that already exist with the aim of making the lives of future generations better off, or at least of ensuring that we allow them to inherit a world at least as good as the one we did. Just as the existence of something like the United Nations can be evidence for the possibility of

---

82 For a general overview of the issues involved, see Kleingeld & Brown (2014).
cosmopolitan commitments across geographical space, the existence of something like global initiatives to fight climate change could be evidence for the possibility of commitments to future generations.

Arguments that can show there are good reasons to say that we have strong obligations to (geographically) distant strangers might also be able to supply principles which could be applied to future generations, provided that the additional issues (like the non-identity problem) could also be addressed. A complete account of such obligations must therefore include empirical and normative arguments to show that the necessary psychological dispositions are both possible and desirable, respectively.

Summary

In this section, my aim has been to describe at least four areas of concern that arise once we start considering our moral obligations to future generations. First, I explained the non-identity problem: it is difficult to figure out the normative relationship between people who are alive now, and people who are not yet alive and whose very lives are contingent on the actions of those who are alive at present. I then showed that one way to avoid this problem might be to adopt an impersonal account of our moral obligations such that we are concerned with, for example, the total amount of utility in the world, rather than the lives of specific individuals. As we have seen, however, this sort of approach leads us to the repugnant conclusion – we should conclude that it is better, for example, to have a world with trillions of people whose lives are barely worth living than it is to have a world with billions of people leading very good lives. Aside from these problems, I also highlighted two other areas that would need to be addressed by a complete theory of justice for future generations. The first of these concerns the various non-normative issues with regard to how we use resources and how we understand what the needs of future generations are likely to be. Secondly, I highlighted the issue of motivational capacities that will be required for present generations to actually act on the obligations that they have towards future generations. Here I drew a parallel with the debate over motivations and cosmopolitan theories of global justice. In the following section, I show how each of these problems might be conceived of using the language of the constraint-based view.
6.3 – Re-Describing the Debate

The preceding section is not intended to be an exhaustive summary of all of the questions that are necessary and sufficient to provide a complete account of our moral obligations to future generations. My intention here has been to highlight some of the most important issues that must necessarily be dealt with, and (as I will demonstrate below) that provide examples of each type of constraint according to the CBV. While I fully expect that other important questions must also be answered before a complete theory of our obligations to future generations can be given, I believe any such questions will be analysable in terms of the CBV as it currently stands. Of course, if it turned out to be the case that there was some particular obstacle which did not fall neatly under one of the existing categories of constraint, additional categories could be added as necessary. For now, it will suffice to stick to the categories as previously outlined.

Recall that the first conceptual distinction in constructing the CBV, is the distinction between A-Type and P-Type approaches. A-Type approaches have two defining characteristics in particular: (1) The intent of the project is to give us a recommendation about what we should think about some particular normative concept(s) and (2) The theory is constrained only by the most basic logical and conceptual constraints. In contrast, P-Type approaches (1) are intended to provide us, ultimately, with action-guiding recommendations and (2) are constrained by hard constraints at a minimum, and may include additional soft constraints, depending on the approach and the case in question.

In the literature on obligations to future generations, it is not always clear which type of approach is being used, and there have been some disputes that may reflect this lack of clarity. So, for example, James Woodward makes the following remarks, while discussing Parfit’s views on the non-identity problem and related issues:

My second reason for dissatisfaction with Parfit’s procedure here is that it seems hard to deny that what is ultimately of interest- the point of the whole exercise, if it is not to be purely academic- is the evaluation of possible actions, policies, and institutional arrangements. But Parfit’s discussion strongly suggests that such real life cases will often have (potentially) morally relevant features that are excluded from most of the hypothetical cases that motivate his discussion of [the repugnant conclusion and [the mere addition paradox]. (It is precisely because such features are potentially relevant that Parfit must take such care to exclude them from many of his hypothetical cases.) Even if Parfit were to find his unknown principle X which would yield an adequate treatment of the RC and so forth, we would still face the complex additional task of determining which (if any) other moral considerations were relevant in real life cases, determining how these should be integrated with or weighed against the purely consequentialist considerations embodied in X, and of assessing the implications of all of this for action. Given that we will want to do this
at some point anyway, it is not clear how much is gained by focusing (largely) on cases from which such considerations have been systematically excluded by stipulation (Woodward, 1986, p. 831).

This kind of objection bears notable similarities to the general thrust of Colin Farrelly's objections to Rawls, and especially Cohen, which I discussed in Chapter 2 (though it is by no means as common an objection in this context as it is in the debate over ideal and non-ideal theory). The debate over obligations to future generations probably lends itself to this kind of confusion, given the way that the debate typically unfolds. By this, I mean that people tend to begin thinking about obligations to future generations because they care about the impact that our actions have on the real world and they want to know how to behave ethically towards future generations. This is a distinctly P-Type enterprise, since the ultimate concern is with producing action-guiding recommendations for the real world, and since we want to (ultimately) take all relevant moral considerations into account (via taking all relevant constraints into account). However, as I discussed in the first section of this chapter, once we begin to investigate these questions with regard to obligations to future generations, we uncover deeper, more abstract problems, like the non-identity problem and the repugnant conclusion, which need to be answered before we can move back towards the real world in deciding what we ought to do. Thus, one could begin by being concerned with the P-Type questions about what we ought to do to meet our obligations to future generations, before being 'pushed back' to A-Type questions concerning the structure of obligations to future generations in the more abstract sense in order to then move back to the P-Type enterprise, on firmer theoretical foundations.

Bringing the A-Type/P-Type distinction to bear on this debate can, therefore, allow for a greater degree of clarity, by allowing us to precisely distinguish between attempts to figure out what we should think about the structure of and relationship between abstract concepts that form the foundation of real-world recommendations, and attempts to produce those recommendations themselves. I should note at this point that the A-Type/P-Type distinction does not require that theorists only engage in one or the other, during the course of a single project, or even a single paper. If I am right that many theorists writing about future generations end up engaged in A-Type exercises because they have been ‘pushed back’ from the initial P-Type questions, then it is clear that many theorists will want to ultimately combine both approaches, or develop both approaches simultaneously. Nevertheless, acknowledging that there are two distinct, if often related, types of questions to be answered
should be useful in clarifying the issues at hand, and in forestalling mistaken objections based on conflating the two approaches.

In the context of questions about obligations to future generations, examples of A-Type questions could include questions like “is the person-affecting principle true?”, “can we benefit a person by bringing him or her into existence?”, “how should we understand the concept of 'harm'?”, “how can we avoid the repugnant conclusion?” and so on. These questions will tell us more about the conceptual structure of moral obligations more generally, before we apply our conclusions to situations in the real world. Provided that our conceptual view ends up being sufficiently robust, applying the general theory that results to particular cases will then yield conclusions about how we may and may not act towards future generations. These conclusions will amount to an articulation of the relevant moral constraints that must be taken into account when we go on to produce action-guiding recommendations as part of a P-Type theory. Thus we can see that, though the two approaches have distinct aims, A-Type theorising is likely to be vital in helping us to construct P-Type approaches that are (and must be) sensitive to moral constraints.

Since (I am assuming) moral constraints are always hard constraints, they will form the foundation for any subsequent P-Type view, regardless of where it sits on the (multidimensional) spectrum of more and less ideal approaches to the issues in question.

Of course, as noted earlier in this chapter, we don't yet have a consensus on a range of important conceptual questions, the answers to which will determine the content of the moral constraints that we must take into account when trying to produce action-guiding recommendations for the real world. The constraint-based view can take this into account, however, by identifying the current disagreement about the structure, content and scope of moral constraints as being a set of constraints upon our behaviour, which any P-Type theory must acknowledge. In other words, any theory about obligations to future generations which seeks to tell us what to do in 2015 is going to be constrained by the fact that we disagree as to what moral constraints apply, because we disagree as to the right answers to A-Type questions. This disagreement can be thought of as constituting a set of epistemic constraints if we believe that we don't yet have enough information to settle the important conceptual questions (many theorists, for instance, believe that nobody has yet solved the non-identity problem or given a satisfactory answer to the repugnant conclusion).
The CBV can respond to this form of epistemic uncertainty in different ways, depending on our analysis of the epistemic constraints in question. In some cases, for example, we will be sufficiently certain about our position to the extent that we will have identified a small range of views, one of which we are certain corresponds to the correct view about, for example, the moral facts which should constrain our actions. In these cases, we can experiment with the different views in question, taking each view in turn and working out the implications of that view on the assumption that it turns out to be true. Given our uncertainty as to which view is actually correct, we would then consider the risks associated with each option, given that whichever one we pick might turn out to be the wrong choice. This analysis would incorporate considerations about flexibility and path dependency (an option that leaves other paths open is more valuable than an option that closes them off in this regard) as well as being sensitive to the different degrees of confidence we might have in the truth of each available option.

In other cases, the epistemic uncertainty may be so deep that we are unable to derive a complete list of candidate views with which to experiment (or we may lack the capacity to tell whether the views we are considering represent a complete list of the views we ought to consider). Given that the epistemic constraints are more powerful in such cases, we should expect that the prospects of coming up with useful, action-guiding recommendations will be relatively slim. It seems to me that what the CBV would recommend in such cases is that we try to identify the capacities that would help us construct a complete list of candidate views, and then that we try develop those capacities, if possible, in order to weaken the epistemic constraints that apply. If these constraints are judged to be hard constraints, however, then it would not be possible to develop the necessary capacities, in which case we would need to apply some more general theory about how to make decisions in cases of moral uncertainty.\(^3\)

On the other hand, if a theorist believes that they do know all of the relevant facts with regard to moral constraints, then they may choose to regard the existing disagreement as a kind of physical constraint in the sense that others don't have physical access to the necessary information, or that we are simply not smart enough to use the information we have to figure out the right answers, which can itself create motivational constraints that inhibit the kinds of actions we would otherwise want to perform (assuming that the theorist's broader view is correct).

There are, of course, other kinds of epistemic constraints that must play a role in any action-guiding recommendations about our obligations to future generations. These constraints concern our capacity to estimate accurately the interests of future generations. If we suppose, for example, that there is some particular account of our moral obligations such that we are obliged to ensure that the lives of future people go sufficiently well, it would remain to be seen precisely how we might discharge such an obligation, given a lack of information about what interests future people are likely to have.

There seem to be two broad types of cases that would be likely to arise. First, there may be needs that future people will have, but which we cannot currently predict that they will have. This inability to predict a future need may represent a hard or a soft constraint, depending on the case in question. Suppose, for example, that at some point in the distant future, a new technology is developed, that allows humans to live and travel in space, but that this technology requires large amounts of some particular resource. In the future, let's say that having access to enough of this resource tends to be necessary for human beings to live fully flourishing lives. At present, we cannot predict what this resource might be, and this fact acts as a hard constraint when we are considering what we might owe to future people, especially those who will live in the distant future. There may be some point in the less distant future, where people might be able to predict that the resource will eventually become necessary. At that point in time, the epistemic constraint moves from a hard constraint to a soft constraint, since it is then possible for people to develop the necessary capacities to predict the emergence of a new need that will affect subsequent generations.

A second type of case concerns needs that we assume will apply to future people, but where we lack the power to predict how best to affect those needs. For example, we can assume that among the strongest interests of any future generation will be an interest in living in an environment with safe, or healthy levels of pollution.\(^84\) While we can assume that these interests exist, it may be virtually impossible for us to know how best to ensure this, given the limitations of our ability to predict the impact of current human activities for future climate change, for example. Sometimes, these limitations will be soft constraints – by researching the effects of human activity on global climate change, for example, we can produce better recommendations about how to minimise pollution or adverse weather conditions for future

\(^84\) Presumably some degree of pollution will be inevitable but worth the cost – it would be better to live in a world with some pollution, but with cars and planes, for example, than to live in a world entirely free of pollution, but where people’s standard of living was lower as a result.
generations. Sometimes these may represent hard constraints, however, especially if we are considering distant future generations and if we have developed out predictive capacities as much as possible. Of course, decisions about what and how to research will themselves involve costs, given that researching ways to tackle any particular problem via any particular method will involve resources that might have been used to research other problems, or other ways of solving the problem in question.

What does seem clear is that a complete theory of our obligations to future generations that is capable of producing action-guiding recommendations will need to incorporate a broader account of decision-making under conditions of uncertainty, given the inevitable presence of both hard and soft epistemic constraints of the sort outlined above.

Provided that we have access to the right moral theory of obligations to future generations (such that we have managed to identify all of the relevant moral constraints) and that we have got a theory that can tell us how to handle the epistemic constraints mentioned above, the CBV allows us to identify a further two categories of constraints that will need to be tackled.

The first of these are physical constraints, represented by the fact that we have limited physical resources that must be divided such that some of our physical resources are distributed to those presently living, while some is saved or invested for the benefit of future generations. The assumption of scarcity of resources can safely be labelled as a hard constraint, provided that saving or investing for the benefit of future generations always represents a cost to present generations (which seems like a reasonable assumption). Presumably, we need a broader theory of distributive justice to help us to explain precisely how resources ought to be divided between generations. Given the uniquely asymmetrical and non-reciprocal relationship between present and future generations, it may well be the case that whichever theory of distributive justice we opt for will require some modification.

While theorists who investigate obligations to future generations are primarily motivated by the worry that we might fail in our obligations to future people, there is obviously a similar danger that if we set too much aside for future generations, we risk failing in our duties to our contemporaries instead.

---

85 However, there may be cases where saving for future generations provides benefits to those of us in the present, even as it imposes costs in terms of limiting some resources. For example, slowing the rate of growth might sometimes have stabilising effects on an economy, reducing the impact on people from cycles of boom and bust.
Assuming we have got the correct account of the moral relationship between current and future generations (by accounting for all relevant moral constraints), and assuming that we have got the right account of distributive justice, and decision-making under conditions of uncertainty (accounting for physical and epistemic constraints respectively), people may still fail to discharge their duties to future generations owing to what the CBV categorises as 'motivational' constraints.

These constraints represent the psychological difficulties that contemporaries may have in becoming sufficiently motivated to do whatever it is that a theory requires of them. Such difficulties are well-known to theorists dealing with global justice and our moral obligations to geographically distant strangers. In the case of obligations to future generations, the distance is mainly temporal rather than geographic (though of course, it could be both). Perhaps work on geographically distant strangers might serve as the foundation for recommendations aimed at minimising or eliminating motivational constraints with regard to temporally distant strangers, though it won't necessarily be the case that conclusions about one type of relationship can be applied without modification to questions about the other.

The extent to which such constraints should be considered hard or soft is a highly controversial issue, and it is clear that any account of obligations to future generations will be shaped by what the particular account assumes with regard to the limits of human motivational capacity. For example, a ‘pessimistic’ theory might assume that human beings are under hard constraints that severely limit their ability to motivate themselves to do what justice might otherwise demand. A more ‘optimistic’ theory, in contrast, might regard such facts as soft constraints, which could be overcome provided that we take appropriate steps to increase our capacity to motivate ourselves to discharge our obligations to future generations.

---

86 See, for example Birnbacher (2009) and Bourban (2014).
87 To be clear, these psychological constraints are those which apply in cases where people agree that the correct moral theory generates obligations to distant strangers, yet where they still find it difficult or impossible to adhere to these obligations. Such people are distinct from those who deny that the correct moral theory actually generates such obligations at all. It is important for a constraint-based view to treat these cases separately, given that the type of constraint is different in each case (epistemic in the case of those who disagree about what morality requires, and psychological in the case of those who find it difficult to do what they believe morality requires of them).
6.4 The advantages of the Constraint-Based View

In the previous section, I offered an account of how we might re-describe debates about obligations to future generations, using the language of the CBV. I take it that this alone is clearly not sufficient to show that we ought to think about these kinds of questions in terms of the CBV. For that to be the case, the CBV needs to offer distinct advantages. In this section, I will explain what I take those advantages to be.

Firstly, using the CBV will help to ensure clarity with regard to the questions under consideration. Specifically, the distinction between A-Type and P-Type approaches allows theorists to be explicit with regard to the different sorts of assumptions they will be using as the ‘inputs’ to their theories, and with regard to the ‘outputs’ that their approach is ultimately intended to produce, whether these are recommendations about what we ought to think or what we ought to do. In addition to clarity about these issues, the CBV allows theorists to further describe each particular assumption in terms of which category it belongs to (moral, physical, epistemic, or motivational) and in terms of its strength – i.e. whether it is a hard or soft constraint and, if it is a soft constraint, how soft it is.

This is not to say that clarity on these matters can only come via the CBV – the claim here is that the CBV merely provides a sufficiently clear way to view these questions, not that it is necessary in order for these questions to be clearly expressed. In this sense, the CBV offers an advantage, but not necessarily a distinct advantage. However, the clarity offered by the CBV becomes particularly important once we move from explaining the questions we are concerned with, to trying to provide answers to those questions. By conceptualising the issues at hand in terms of hard and soft constraints in the context of a P-Type approach, we can make use of both the “narrowing-down approach” and the “transitional principles” described previously in Chapters 4 and 5.

Recall that the narrowing-down approach requires us to take steps to improve our capacity to identify relevant constraints, and to measure the strength of those constraints. This strategy has clear implications for the debate on obligations to future generations. For example, it explains why it is important to take steps to ensure that we can accurately predict the effects of our actions on the interests of future generations (thereby ‘softening’ existing epistemic

89 However, approaches which fail to make similar distinctions as between hard and soft constraints, or a-type and p-type approaches are likely to be less clear in general than those, like the CBV, which do, for the reasons outlined in Chapter 3.
constraints that prevent us from knowing how best to respond to the needs of future generations). It also explains why it is important to investigate the limits of human motivational capacities in general, in order to ensure that our theories make the correct assumptions about motivational constraints. Most obviously perhaps, the narrowing-down approach explains why it is important to do A-Type theorising in addition to P-Type theorising about obligations to future generations, since A-Type theorising involves the identification of the values which will form part of the moral constraints that need to be incorporated into P-Type recommendations.

Finally, consider the ‘transitional principles’ sketched in Chapter 5. Having used the CBV to help identify the ‘ideal’ (i.e. a world where only hard constraints apply and we have met all of our obligations to future generations to the greatest possible extent), applying these principles to questions of obligations to future generations can provide us with a helpful way of identifying circumstances where deviation from pursuit of this ideal is permissible. These could include, for example, cases where pursuit of the ideal cannot be achieved except by violating what I called ‘deep moral constraints’, or where pursuit of the ideal cannot be achieved except by violating a public reason condition, or where it would be too demanding, or highly infeasible, and so on, depending on which particular transitional principles we accept. In identifying the circumstances under which deviation from the pursuit of the ideal is at least permissible, we can obviously also identify cases where pursuit of the ideal is required, in cases where we determine that none of the relevant ‘excuses’ for deviation apply.

6.5 A non-ideal approach to the non-identity problem

Having explained how the language of the CBV can be used to re-describe some of the debates surrounding obligations to future generations in Section 6.2, and having shown in the preceding Section that thinking about these questions in terms of the CBV provides us with distinct advantages, my aim in the final section of this chapter is to sketch an example of how one of these issues might be approached with the help of the CBV.

Before I begin, two points of clarification are required. First, my claim is not that the CBV is the only effective way to approach these issues, but merely that it is a particularly effective way (as I will hopefully be able to demonstrate below). As I suggested in the previous section, an approach that is not based on the CBV may be perfectly capable of providing us
with action-guiding recommendations that are achievable and desirable, but it may also lack
the distinct advantages of the CBV in terms of providing a clear account of all of the relevant
constraints and recommendations that are specified in terms of dealing with particular
constraints. Second, the particular solution I will describe in this section is not a view that can
only be arrived at via the CBV. There are several authors who have considered broadly
similar arguments based around the idea of overlapping generations. However, the CBV
does provide a clear and straightforward path towards the solution in question, and identifies
it as a distinctly non-ideal approach – a characterisation that is often not present in similar
arguments. With these caveats in mind, I turn now to the question of how the CBV might
help us to tackle some of these issues.

One problem that is central to many of the questions surrounding future generations can
be characterised as follows: we have strong intuitions to the effect that we have moral
obligations to future generations, but we seem to lack a compelling account as to how
precisely these obligations are generated. Here are two ways in which our intuitions seem to
lack convincing philosophical foundations:

The Person-Affecting Problem: Many people have the intuition that if something is bad,
then it has got to be bad for someone. However, future generations don’t yet exist, so if we do
something now, which will only cause harm to people who will not exist until long after we
are dead, and whose very existence is contingent on whether we perform the act in question,
how can it be that we have done something bad?

The Quality of Life Question: This question concerns whether we have an obligation to
create future people who will lead optimal lives or whether it’s permissible for us to create
future people of any sort, provided that their lives go sufficiently well. If we say that we have
no such obligation, then we may have to accept some counterintuitive claims (like the claim
that it’s okay for a couple to take a ‘pleasure pill’ before conception which they know will
result in them producing a disabled child whose life is nevertheless worth living). If we say
that we do have an obligation, then we may have to accept different but equally

90 Axel Gosseries (2008) identifies an approach which he calls the “present rights of present persons” option
which is the type of solution I will argue for in this section. Richard Howarth (1992) describes what he calls a
“chain of obligation” between overlapping generations as part of a view which is very similar to the solution
described in this section (though there are some significant differences, such as my account of the precise
content of our obligations to present generations and the distinctly non-ideal flavour of the solution as I describe
it). See also, Steiner & Vallentyne (2009) on the issue of overlapping generations and conceptual theories of
rights and Gardiner (2003) on why a form of the ‘Pure Intergenerational Problem ‘persists even in cases
involving overlapping generations.
counterintuitive claims (like the claim that a woman who chooses to have a child at a young age does something wrong by not waiting to have a child at a later point provided that doing so would produce a child with a better quality of life).

An ideal solution will provide us with convincing answers to both of these problems. It will provide an account that can explain either why the person-affecting principle can be extended to cover future generations, or why obligations to future generations can be sustained even in the absence of the person-affecting principle. The ideal solution will also give us a convincing explanation of the content of our obligations with regard to creating future generations, and ensuring they receive a particular quality of life.

Applying the Constraint-Based View to these problems suggests that we focus on identifying the kinds of constraints that prevent us from achieving an ideal solution at present. While the exact description of the constraints will depend on the theorist that is describing them, the main obstacles to an ideal solution can be viewed as a series of epistemic constraints surrounding either a lack of knowledge about the relevant moral facts or a lack of agreement about the relevant moral facts. Of course, even once these constraints are removed, other constraints may well present themselves – if we decide that we do have obligations to future generations, for example, it is likely that many more epistemic and perhaps motivational constraints will need to be tackled in order for us to fully articulate and discharge our moral obligations.

In order to explain what a non-ideal approach to these problems might look like, I want to introduce the idea of what I will call “moral over-determination”. An action is morally over-determined for some particular moral agent if it is the case that the agent has more than one decisive reason to perform the action, where a ‘decisive’ reason is one that is sufficient in the context in question to generate a duty to perform the action. For example, suppose a child promises each of her parents separately that she will do her homework. Promising either parent alone would have been sufficient to generate the duty to complete the homework, so the action is over-determined in this case. Note that for an action-to be overdetermined the strength of each reason need not be equal, as in the homework case. For example, suppose Alice is contemplating whether to hit Betty. Alice knows that Betty has a right to bodily integrity and Alice knows that assaulting Betty would make Betty’s friend Charlie extremely upset. In this case, the right to bodily integrity is far weightier than Charlie’s interests in not
being upset, but the latter would have been sufficient on its own to count against Alice hitting Betty.

Now suppose that Alice is contemplating whether to hit Betty, and that she has the strong intuition that it would be wrong of her to do so, but she isn’t sure of how to explain why it would be wrong, because (for some strange reason) she doesn’t really understand the idea of a right to bodily integrity. In contrast, Alice is very certain that it would be wrong of her to upset Charlie (a moral belief that rests on something other than a right to bodily integrity, obviously) and so she decides not to hit Betty, because she is sufficiently certain of one of the decisive reasons she has.\(^9\)

In one sense, Alice’s concern for Charlie’s feelings can be seen as a ‘non-ideal’ approach to the problem in the sense that Alice arrives at the desired outcome (discharging moral duties) via an appeal to reasons, which, while sufficient to generate moral duties, are not necessarily the weightiest reasons she might have for acting. In a sense, the epistemic constraints that prevent access to these weightier moral reasons have been side-stepped – allowing Alice to arrive at the right destination, even if she hasn’t necessarily taken the route that provides her with the weightiest moral reasons.

It is important to emphasise in these each of cases that the weaker reasons are sufficient to generate obligations to act. It may not be satisfying to act on those reasons, because our intuitions are that there are better reasons out there, but at no point are we required to hold false beliefs, or take a risk of acting wrongly. We have sufficient reasons to act in each case; we just don’t have the best reasons, and ideally we would.

Here is what a similar approach might look like in the context of obligations to future generations, given the person-affecting problem and the quality of life question:

Suppose that we have a set consisting of all the possible children to which some particular mother might give birth over the course of her life. Within this set, we can identify a subset of possible children who will lead lives worth living, and within that subset we can identify a

\(^9\) If this example seems too strange, here is another: some people are uncertain whether animals have moral status, and so they are uncertain whether animals have a right not to be tortured for fun. When pressed as to whether it should be illegal to torture animals for fun, such people may instead appeal to the idea that torturing animals should be illegal because it will make animal-torturers more likely to violate the rights of human beings in the future. The plausibility of either claim is not important for the purposes of this example, which is simply intended to illustrate a case where a weaker, but more certain moral reason allows us to perform an action (in this case, prohibiting animal cruelty) even though there may be a much stronger reason to perform the same action (if animals do have moral status, this looks like a much weightier consideration).
further subset of possible children who will lead lives that are not just worth living, but will have a good chance of living a truly flourishing life. Call these ‘optimal children’. It seems plausible to me to say that the mother has a strong interest in being able to choose to have an optimal child. To support this claim, consider the case of the girl who is deciding whether to have a child now, at the age of sixteen when she has few resources available, or to wait until she is older and will likely be in a better position to provide a good life for her child. Suppose that if she has a child now, she knows it will have a life worth living, but won’t be an optimal child, but if she waits she can have an optimal child. It seems clear to me that if someone were to intervene to remove the option to wait, that they would have damaged the girl’s interests in an important way.

Note that we can agree with this while remaining entirely agnostic as to whether it is permissible for the girl to have the child now, or whether she is required to wait.\textsuperscript{92} The interest in question here is not to be confused with the interest each of us has in ensuring that our loved ones live worthwhile lives (since no child yet exists at this point) but rather an interest in being able to create and sustain worthwhile relationships in the future that may not exist at present. This interest is grounded in both the potential benefits that such relationships may bring, as well as the benefit of being able to autonomously choose what sort of relationships to create.

Let’s call the capacity to choose to have an optimal child, Capacity X. Given the strong interest the mother has in having Capacity X, we can say that it will be more difficult for the mother, and people like her, to lead a flourishing life in cases where they lack Capacity X.

So, if we alter a potential (male or female) parent’s range of options, such that we deprive them of Capacity X, we harm them to the extent that we make it less likely that they will be able to live a flourishing life.\textsuperscript{93} If we remove the girl’s option to wait, for instance, this constitutes a harm to her, even if she would not have chosen to wait, and even if she would not have chosen to have a child, provided that having that option open to her is valuable in

\textsuperscript{92} One may have the intuition that it is not merely permissible but required for the woman to wait, in such a scenario. Though I do not share that intuition, my argument here depends only on the assumption that it is at least permissible for the woman to wait.

\textsuperscript{93} The harms caused by damaging a person’s capacity to create optimal children may sometimes be offset by benefits which are provided as a result of damaging that capacity. This fact does not pose a significant challenge for my view, however since it suggests merely that the interests of present generations may sometimes permit a failure to create optimal future generations.
itself. Given this, it looks like we have (at least) a prima facie reason not to hinder a person’s capacity to choose to create optimal children.\textsuperscript{94}

Furthermore, we can say that a person who lacks Capacity X, will themselves be less likely to lead a flourishing life of their own, since they will have been denied a choice that they have strong interests in having. In other words, being an optimal child requires the capacity to create optimal children, if one chooses. It should be emphasised again that I am not assuming that one cannot live a flourishing life without choosing to create optimal children, or even without possessing the capacity to create optimal children. What matters here is the interests people have in having significant choices, like the choice to create an optimal child, which will make it more likely that a person lives a fully flourishing life, because those options are open to them.

To summarise the above, here are four claims: (1) People have a strong interest in being able to choose to create optimal children [Capacity X]. (2) People who lack Capacity X will be less likely to live flourishing lives. (3) We have a prima facie reason not to hinder a person’s possession of Capacity X. (4) Being an optimal child requires the possession of Capacity X.

Given these claims, suppose we have 26 generations of people, from Generation A to Generation Z. Suppose that Generation A adopts a policy that will see a rise in the standard of living for each subsequent generation until Generation Y, at which point members of Generation Y will only be able to create sub-optimal children (children who lead lives worth living, but whose lives are not likely to be flourishing lives).

It follows from claims 1-4 above, that Generation Y lacked Capacity X, which means (following the fourth claim) that the members of Generation Y are themselves sub-optimal. If the members of Generation Y are sub-optimal, and if the capacity to create optimal children is necessary for one to lead an optimal life, then it follows that Generation X are sub-optimal. Thus we have a regression, which takes us all the way back to the actions of Generation A, which renders the members of Generation B sub-optimal, thereby violating the duty upon Generation A as per the third claim, above.

\textsuperscript{94} This reason is not without its limitations, of course – an absolute commitment to maximizing a person’s capacity to create optimal children would likely place too much emphasis on the interests of future generations compared to those presently living.
Now recall the two problems outlined previously: the person-affecting problem and the quality of life question. The person-affecting problem consists of the fact that we think that in order for someone to do wrong, it must be the case that their actions have actually harmed the interests of another person who actually exists. The story I’ve told above, however, can satisfy the requirement that any wrongs we identify must be able to point to persons who actually exist, since there is an overlap in time between the members of Generations A and B – because members of Generation B live at the same time as members of Generation A, we can point to them in explaining why the actions of Generation A wrongs people who are actually living at the point in time when the initial decision is made.

Now, intuitively, we don’t just think that the members of Generation B are being wronged. We want to be able to say that everyone from Generations B to Z is being wronged, and especially members of Generation Z, since they will have worse lives than the preceding generations. Unfortunately, a plausible and persuasive explanation for such a conclusion is not available at present [and this fact constitutes an epistemic constraint upon our behaviour]. Ideally, we would have such an explanation, but for now we can at least point to the wrongs committed against the members of Generation B in order to explain why the policy should not be adopted by Generation A.

Turning to the quality of life question, recall that this problem concerns the question of whether it is permissible to create sub-optimal children who nevertheless lead lives worth living, or whether there is an obligation upon people who choose to create children, that they create optimal children where possible. The upshot of the non-ideal approach is that it avoids this question, since the non-ideal approach depends only on the assumption that people ought to have the choice to create optimal children, rather than assuming that they ought to make that choice where available. Again, ideally we would have access to a plausible and persuasive answer to the question, but in the meantime at least we can explain one of the reasons why Generation A’s members are acting impermissibly in adopting the policy in question.

Following an approach like this allows us to get closer to the point where we can issue action-guiding recommendations for the real world, by side-stepping some of the epistemic constraints that would otherwise have prevented us from rendering a judgment about the permissibility of our actions. Of course, the approach I’ve just described only focuses on some of the constraints that would need to be eliminated. A fully-realised theory will likely
require input from experts in various fields – presumably scientists and engineers will be able to give us ways to solve many of the relevant physical and epistemic constraints, psychologists might be able to help us solve any relevant motivational constraints and so on. As the process that produces action-guiding recommendations nears completion, I imagine that the work of philosophers will be confined to tackling the sorts of constraints represented by the person-affected problem and the quality of life question.

One might object that the non-ideal solution I’ve proposed in this chapter is fine as it applies to obligations to future generations, but that we have little or no general guidance when faced with similar problems elsewhere. That is to say, it is not clear (so the objection might go) how exactly it is that the CBV helps us arrive at this solution, and whether it would be likely to help us arrive at similar solutions when dealing with other cases.

It is important when responding to this sort of objection not to ‘over sell’ the usefulness of the CBV. Adopting the CBV should usually be helpful in allowing us to come up with a clear description of what the problems are, which we can express in terms of the various constraints that stand in the way of issuing action-guiding recommendations. However, while this process may lead to a clearer understanding of the problem, there is no guarantee that it will lead us directly to the solution (if a solution exists). Additionally, the solutions recommended by adopting the CBV may be arrived at without using the CBV – as I noted earlier, there are other approaches to obligations to future generations that are similar in many ways to the one outlined in this chapter. One advantage of using the CBV, however, is that it identifies the solution as a distinctly non-ideal approach to the problem, and identifies the constraints that would need to be eliminated to arrive at a complete theory of our obligations to future generations.

In this case, the CBV is helpful because its focus on constraints leads us to the conclusion that we need a way to respond to the epistemic constraints surrounding disagreement over things like the non-identity problem and that we should see how far we can move towards an ideal solution in the meantime, given the constraints that we believe will be in place for the foreseeable future. In this case in particular, it looks as though we can move forward via an appeal to the relatively weak (but still important) moral considerations outlined above (i.e. the interests presently living people have in being able to create optimal children). Whether

---

95 See Herzog (2012).
similar routes are possible in other contexts will obviously depend on those contexts, but the CBV can at least give us a clearer idea of what such solutions would look like.

In a sense, adopting the CBV may itself constitute part of the very “narrowing down” process it recommends. By adopting the CBV we may improve our capacities to identify and assess relevant constraints, and formulate political proposals aimed at moving us from the status quo, towards a more ideal world.

6.6 Conclusion

My aim in this chapter has been to offer an example of the Constraint-Based View in action. The example I chose, that of intergenerational justice, is but one of a great many of cases to which the CBV might apply, and I have only been able to describe how the CBV might help us to tackle some of the constraints that will be relevant when producing a complete theory of intergenerational justice. What I have sought to show in this chapter is that it is possible to translate debates about intergenerational justice to the language of the CBV, and that thinking about the problems in this way offers some unique advantages. Specifically, the process of identifying and describing the relevant constraints, and of plotting a path towards a more ideal destination (in this case, where we can begin to issue recommendations, even if we lack access to the best reasons for our actions) should help us tackle these and similar issues in a way which is clear, consistent, and offers a real prospect of moving from the status quo, to a more achievable and desirable state of affairs.
Conclusion

My primary aim in this thesis has been to argue that the Constraint-Based View is the best way to think about the relationship between ideal and non-ideal theory. In Chapter 1, I began my case for the CBV by assessing some of the ways that theorists have attempted to draw a distinction between ideal and non-ideal theory. In some cases, these distinctions were overly narrow – capturing only one of the many ways in which a theory can be more or less ideal (such as the distinction between full and partial compliance). In other cases, the distinctions did not seem to plausibly map on to ideal and non-ideal theory (such as Sen’s distinction between ‘transcendental’ and ‘comparative’ approaches to justice or Farrelly’s distinction between sufficiently fact-sensitive and insufficiently fact-sensitive approaches). In other cases, the proposed distinction was simply too unclear to be useful (Mills’ distinction between idealization and abstraction). Despite rejecting these conceptions, the CBV turned out to be compatible with Hamlin & Stemplowska’s conception of a multidimensional continuum of theories that are more or less ideal in multiple respects.

Before offering a description of the CBV itself, I described and defended a distinction between two different theoretical approaches to justice in Chapter 2. The purpose of this chapter was to allow me to locate the CBV within the context of P-Type, rather than A-Type approaches to justice, while explaining why G.A. Cohen ought to be considered an A-Type theorist (and therefore, why much of the criticism levelled against him and similar theorists by Colin Farrelly misses the mark).

In Chapter 3 I described and defended the CBV itself, arguing that we ought to conceive of a theory as being more ideal to the extent that it does not include assumptions about soft constraints as part of its inputs, and more non-ideal to the extent that it does. This view suggests that we think in terms of a spectrum of possible approaches, with ‘hard constraints only’ views occupying the extreme ideal end of the spectrum and views that include every relevant hard and soft constraint at the extreme non-ideal end. I suggested that we categorize these constraints as logical, physical, epistemic, motivational and moral. While I maintain that this way of categorizing constraints is
helpful, it is by no means necessary to accept these particular categories in order to endorse the CBV.

Applying the CBV requires that we are able to identify relevant constraints, assess the status and strength of those constraints, and formulate plans of action that are capable of dealing with a sufficient number of constraints. In Chapter 4, I examined the role that feasibility assessments play when trying to devise action-guiding recommendations. Here I endorsed a view similar to that offered by Pablo Gilabert and Holly Lawford-Smith (which also uses constraint-based language) and focused in particular on our capacity to make accurate feasibility assessments. I described a ‘narrowing-down approach’ whereby we try to improve our capacities to make accurate feasibility assessments in order to rule out obviously inferior political proposals.

In Chapter 5, I focused on those principles that could justify a deviation from the pursuit of an ideal in favour of a next-best alternative. I identified several principles which should guide us in making ‘route comparisons’ and ‘destination comparisons’ and emphasised, again, the importance of developing the capacities necessary to make reliable judgments about political proposals and states of affairs.

Chapter 6 offered an example of what it would look like to apply the CBV to a particular debate in contemporary political philosophy. In this chapter, I described several key questions concerning the possibility and nature of our moral obligations to future generations, and showed how these questions could be framed using the language of the CBV. I then sketched one way in which the CBV could help us approach these problems, by formulating a proposal which bypasses important and (for the moment) persistent epistemic constraints concerning moral facts about our obligations to future generations.

Taken as a whole, I hope this thesis has offered a plausible and persuasive case in favour of a constraint-based view of ideal and non-ideal theory.
Bibliography


