A CONTRACTUALIST THEORY OF NONIDEAL JUSTICE

Matthew Robert Adams
London, United Kingdom

B.A. (First Class hons.), University of Dublin Trinity College, 2009

B.Phil., University of Oxford, 2011

A Dissertation presented to the Graduate Faculty
of the University of Virginia in Candidacy for the Degree of
Doctor of Philosophy

Department of Philosophy

May, 2018
A. John Simmons
John Allen Hollingsworth Professor of Philosophy
(Chair)

Talbot Brewer
Professor of Philosophy

Rebecca Stangl
Associate Professor of Philosophy

Colin Bird
Associate Professor of Politics
(External)
It is obvious that between theory and practice there is required, besides, a middle term connecting them and providing a transition from one to the other, no matter how complete a theory may be; for, to a concept of the understanding, which contains a rule, must be added an act of judgment…But even where this natural talent is present there can still be a deficiency in premises, that is, a theory can be incomplete…In such cases it was not the fault of theory if it was of little use in practice, but rather of there having been not enough theory…

# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgments</td>
<td>i</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Chapter 1: Ideal Theory &amp; Nonideal Theory</td>
<td>5</td>
</tr>
<tr>
<td>Chapter 2: Nonideal Principles of Justice</td>
<td>33</td>
</tr>
<tr>
<td>Chapter 3: The Value of Nonideal Principles of Justice</td>
<td>68</td>
</tr>
<tr>
<td>Chapter 4: An Ideological Critique of Nonideal Methodology</td>
<td>100</td>
</tr>
<tr>
<td>Chapter 5: Nonideal Justice, Fairness, &amp; Affirmative Action</td>
<td>133</td>
</tr>
<tr>
<td>Appendix I: Contractualism &amp; Intergenerational Justice</td>
<td>160</td>
</tr>
<tr>
<td>Bibliography</td>
<td>176</td>
</tr>
</tbody>
</table>
Acknowledgements

The germ of this dissertation was a conversation that I had with John Simmons as a first year graduate student. At the time I was taking his course “Rescue, Charity, and Justice.” One of the books assigned was Liam Murphy’s *Moral Demands in Nonideal Theory*: my first extended exposure to nonideal theory. Towards the end of the semester I mentioned that I had been wondering whether the normative content of nonideal theory could be determined using a contractualist framework. John seemed intrigued but uncertain. Our mutual perplexity inspired me to think more about the topic and eventually to write a dissertation on it.

John turned out to be an exceptional advisor. His enthusiasm for the project from the very outset was a constant source of inspiration; his careful guidance and support every step of the way was invaluable. In particular, his feedback on drafts and his own published work on the topic always pressed me to try to express my own ideas more sharply and elegantly. I could not imagine writing this dissertation without his help.

Tal Brewer and Becky Stangl both went way beyond what could reasonably be expected from committee members. They often drew my attention to new and unexpected problems that I had neglected in earlier drafts. In the short term this made things harder; in the long term it helped me to grapple with the problems more deeply. I am also grateful to Colin for agreeing to be my external examiner and for many enjoyable and rewarding discussions over the last few years—often fueled by whiskey—on topics ranging from dignity to ideology.

During the spring semester 2017 Dave Estlund agreed to take me on as a visiting student at Brown University. Our leisurely lunches throughout the semester were fun and intellectually stimulating. I am particularly grateful to Dave for encouraging me to write a paper on ideology that eventually formed chapter 4 of this dissertation.
I have also received helpful comments on portions of the dissertation from the following people: Michael Carroll, Harrison Frye, Luke Hunt, Donncha Mac Cóil, Kristina Meshelski, Ross Mittiga, Andrew Morgan, Kirun Kumar Sankaran, and Margaret Shea. I would also like to particularly thank Donncha for taking on the large task of proof reading the final version with great dedication and rigor.

More generally, I am grateful to the UVa philosophy department. When I arrived as a graduate student six years ago I really did not know what to expect; thankfully the department turned out to be a wonderful academic home. I value how supportive the graduate students are to each other, and how so many of the faculty are obviously driven by what they judge to be philosophically important rather than what is academically trendy. I have learnt a lot over the last five years and a large part of the credit goes to the people who have surrounded me.

Finally, on a more personal note, I count myself extraordinarily fortunate to have such a wonderful group of friends. The following people are worthy of special mention: Jenn Brewer, Matthew Byrne, Michael Carroll, Robert Cummins, Lisa Devlin, Miho Erin, Corin Fox, Katherine Freeman, Cathy Lanigan O’Keeffe, Donncha Mac Cóil, Leisha Marlow, Richard Matthews, Ross Mittiga, Andrew Morgan, Nicholas Rimell, Abdallah Salam, and Margaret Shea. Last—but certainly not least—I thank my immediate family for their love and support: my sister Sophie, my brother Tom, my mother Lesley, and late father Robert.
Introduction

Ideal theory is a conception of how a perfectly just society should be arranged. The orthodox (Rawlsian) position is that ideal theory is an essential part of a theoretical conception of justice: in nonideal conditions we should transition to the full realization of ideal justice, and this transition should be orientated by ideal theory.¹

This orthodoxy has recently been extensively challenged. Some philosophers—most famously Amartya Sen—argue that it is not necessary to get “heated up” about ideal theory in order to determine what we should do in our actual nonideal world.²

One obvious reason why this orthodoxy has been challenged is that in John Rawls’s work (and in the Rawlsian tradition in general), the claim that ideal theory is an essential part of justice has been insufficiently clarified and defended.

Another reason reflects a deeper social pathology. Rawls wrote the first edition of A Theory of Justice during a period of relative optimism. The civil rights movement brought hope that racial injustice—arguably the most endemic form of injustice in the United States—could eventually be overcome. With the exception of a slight rise between 1950 and 1960, wealth and income inequalities gradually decreased from 1910 to 1970.³ In doing so they (at least partially) confirmed the economist Simon Kuznets’s famous prediction that economic

---


inequalities would automatically decrease in advanced stages of capitalism.⁴ Whilst the basic structure of the United States clearly did not conform to ideal principles of Rawlsian justice in the period leading up to the publication of the book in 1971, it was perhaps reasonable to believe that something close to Rawlsian ideal justice could be achieved in a few generations. Crucially also, it was reasonable to hope that this achievement could come about largely by political and economic inertia.

Sadly such reasonable optimism was short lived. The back cover of Rawls’s late work *Justice as Fairness* informs the reader that ‘Rawls is well aware that since the publication of *A Theory of Justice* in 1971 American society has moved further away from the ideas of justice as fairness.’⁵ Since Rawls’s death the impression of such backsliding has intensified. As recent political events dramatically illustrate, racial injustice remains deeply engrained in the United States. Indeed, if anything (at least in certain respects) it is enjoying a perverse renaissance. Wealth and income inequalities have also continuously risen since the 1970s. Essentially, the gulf between our actual nonideal world and the ideal seems increasingly great and insurmountable. Consequently, ideal theorizing seems increasingly futile.

A central aim of my dissertation is to vindicate the orthodox position that ideal theory is an essential part of justice—a part that is crucial for an adequate development of nonideal justice. The central thread of my argument is captured by the prefatory epigram by Kant: the value of ideal theorizing in our actual nonideal world is obscure because the requisite theorizing is incomplete. My innovation of what I term “nonideal principles of justice” is crucial to overcoming this problem of under-theorization: nonideal principles of justice are “idealized” in the sense that they abstract away from considerations of political feasibility, and specify what justice *simpliciter* requires. But they are “nonideal” in the sense that they specify what should be done in nonideal conditions, rather than specifying how a perfectly just society should be structured. I argue that they are the key to clarifying how ideals of perfect justice relate to what we should do in our actual nonideal world.

In this dissertation I also attempt to confront the deeper social pathology that has propelled suspicion of ideal theory. In one sense this pathology is reasonable: ideal justice is probably not something that will ever be realized in the actual world. Tragically, something

---


like nuclear armageddon seems far more plausible. Where this pathology errs however is in confounding the claim that ideal justice is probably not achievable in the actual world with the claim that it is not instrumentally valuable in the actual world.

I suggest that in order to craft compelling solutions to exigent and contested philosophical problems, such as whether non-meritocratic affirmative action is just, we require a rich theoretical account that can only be supplied by ideal theory and nonideal principles of justice. Furthermore, I contend that in order to adopt the right attitudes to policies that reflect our limitations (e.g., our epistemic limitations and restrictions imposed by political feasibility) we require a conception of justice—supplied by ideal and nonideal principles of justice—that does not reflect such limitations. Even if the gulf between our actual nonideal world and the ideal seems great and insurmountable, we still require the ideal to guide our actions and attitudes.

My dissertation has the following structure: in chapter 1 I survey the contemporary literature on ideal and nonideal theory, and uncover some problems and limitations with Rawls's conception of nonideal theory. I motivate the need to develop Rawls's conception of nonideal theory into a coherent and comprehensive theory of nonideal justice: I argue that this will put us in an adequate position to assess a broadly Rawlsian theory of justice, and will be valuable for even non-Rawlsian work on ideal and nonideal theory.

In chapter 2 I show how a Rawlsian contractualist framework can be used to derive nonideal principles of justice. I argue that these nonideal principles of justice overcome the central problems and limitations with Rawls's position as it stands.

In chapter 3 I illuminate the value of these nonideal principles of justice in actual nonideal conditions. I argue that these principles play, what I term, an institution-guiding regulative function. I also caution that the value of these principles cannot be reduced to this institution-guiding function. They also perform what I term an attitude-guiding function, which facilitates the right response to the institutional realization of justice.

In chapter 4 I confront the charge that ideal theory is ideological: essentially, that it performs the distorting social role of reifying and enforcing unjust features of the status quo. I defuse this charge by showing that an equally plausible ideological critique can be presented against philosophers who reject ideal theory.

Finally, in chapter 5 I show how these nonideal principles of justice can be used to specify the conditions under which non-meritocratic affirmative action policies are just. I
round out this chapter—and indeed my dissertation as a whole—by presenting a more general methodological reflection about the value of the conceptual innovation of nonideal principles of justice.
Chapter 1

IDEAL & NONIDEAL THEORY

1.1 Introduction

In this chapter I lay the foundations for my whole dissertation project. I begin by surveying the various ways in which the distinction between ideal and nonideal theory has been explicated in the literature. I uncover some problematic aspects of this literature, and defend the conception of these terms that I favor.

After that, I outline the sketch of nonideal theory that we receive from John Rawls. I argue that this sketch—at least as it stands—is underdeveloped and faces a number of problems. However, I also highlight the importance of developing this sketch of nonideal theory into a comprehensive theory of nonideal justice: I argue that this will put us in an adequate position to assess a broadly Rawlsian theory of justice, and will be valuable for even non-Rawlsian work on ideal and nonideal theory.

1.2 Overview of the contemporary literature on ideal and nonideal theory

I begin by presenting a synoptic overview of the literature on ideal and nonideal theory.

1.2.1 Characteristic features of ideal and nonideal theory

An ever increasing number of political philosophers have worked either directly on the topic of ideal and nonideal theory, or on topics that intersect with this topic in some way.
In doing so, they have bequeathed a variety of conceptions of the subject matter of ideal and nonideal theory to subsequent scholars. These different conceptions can be condensed, with some simplification, into the following “characteristic features” of ideal theory:

(1) A conception of how a perfectly just society should be structured.¹

(2) Constructed using the method of idealization. For example, principles of justice may be built on an assumption that actual people will strictly comply with the principles selected. This assumption does not, of course, correspond with how actual people will in fact behave.²

(3) A top-down approach: very abstract principles of justice are accorded theoretical primacy; these are then applied to solve concrete problems.³

(4) Normative principles are determined in independence from rigorous social science.⁴

(5) Fact-insensitive: theorizing that is completely independent from contingent empirical facts.⁵


⁵ This is what G. A. Cohen construes as the defining feature of ideal theory. He argues that ‘…principles that reflect facts must, in order to reflect facts, reflect principles [i.e., be grounded in principles] that don’t reflect facts.’ (G. A. Cohen, “Facts and Principles,” Philosophy & Public Affairs 31 (2003): 211-245, p. 214.) Essentially, all fact-sensitive principles must ultimately “bottom-out” in fact-insensitive principles, and it is appropriate to designate such fact-insensitive principles as “ideal.”
In contrast, the following are characteristic features of nonideal theory:

(6) A conception of what ought to be done in conditions that fail to realize perfect justice.

(7) Constructed using the method of abstraction rather than idealization: in a model of social reality some variables are bracketed and put to one side; however, no idealizing assumptions, such as strict compliance, are introduced.6

(8) A bottom up-approach: begins with concrete problems rather than trying to antecedently determine very abstract principles of justice.

(9) Normative principles are informed by rigorous social science.

6 Following O'Neill, *Towards Justice and Virtue*, 40. I must confess, I have found it hard to understand what the metaphor of “abstraction” amounts to in the context of constructing normative theory. O'Neill uses the distinction to argue that Kant’s methodology is superior to that of Rawls: in contrast to Kant, Rawls idealizes features of agency and rationality rather than abstracting sufficiently and relying on the least determinate conceptions of rationality and agency. (Onora O’Neill, *Constructions of Reason: Explorations of Kant’s Practical Philosophy* (Cambridge: Cambridge University Press, 1990), 210.) However, the claim that Kant—in contrast to Rawls—abstracts rather than idealizes is perplexing. Surely, if anything is a philosophical idealization, then Kant’s regulative ideal of a kingdom of ends is an idealization. I think there is a worry that whatever gloss is given to the concept of “abstraction” it will not be sufficient to distinguish ideal and nonideal theory. For instance, if an “abstraction” must just make reference to an actual state of the world and an “idealization” must be completely independent from any actual state, then even paradigm ideal theories can be classified as abstractions. This is because, as Eva Erman and Niklas Moller point out, “…the sought-after states that ideal theory typically aims for, states of equality, nondomination, and so on, are not mere “fantasy sates”; they obtain, if only on a more local scale.” (Eva Erman and Niklas Moller, “Three Failed Charges Against Ideal Theory,” *Social Theory and Practice* 39 (2013): 19-44, p. 41.) Some theorists, such as Laura Valentini, have tried to make the distinction between abstraction and idealization clearer by arguing that idealizations—in contrast to abstractions—are designed under idealized, i.e., false assumptions. (See: Laura Valentini, “On the Apparent Paradox of Ideal Theory,” *Journal of Political Philosophy* 17 (2009): 332-355, p. 338.) It is, however, wrong to describe idealizations as falsehoods. Indeed, as Holly Lawford-Smith argues idealizations such as the original position can be formulated as truth-apt counterfactuals. (See: Holly Lawford-Smith,” Debate: Ideal Theory—A Reply to Valentini,” *Journal of Political Philosophy* 18 (2010): 357-368.) It is true that some of the antecedents in these idealization (e.g., Rawls’s methodological assumption of strict compliance) would be falsehoods if they were offered as descriptive claims about the actual world. However, this does not mean that the ideal model, itself, is designed under false assumptions. Perhaps the best way of explicating the distinction is to argue that idealizations are used to design a philosophical model (e.g., a contractualist framework), whereas abstractions are intuitions that ordinary folk have outside such a model.
Fact-sensitive: theorizing is contingent on and relative to at least some empirical facts.

These characteristic features bring together the work of different philosophers. Not all theories that, for instance, could reasonably be classified as “ideal” will have all of features (1)-(5). Martha Nussbaum’s capability approach, for instance, satisfies (1) because it treats the achievement of ten central human capabilities as a necessary condition for realizing perfect justice. Her theory does not, however, satisfy (2) because—in contrast to Rawls’s theory of justice—it is not built on any idealizing assumptions, such as strict compliance. Relatedly, some theories satisfy some characteristic features of both ideal and nonideal theory. For instance, Sharon Dolovich’s theory of punishment satisfies condition (6) because it is a theory for a liberal democracy that is not fully just. However, it also satisfies condition (2) because it is constructed using a Rawlsian contractualist framework and, therefore, uses the method of idealization. More remarkably, Rawls’s theory of justice—which is standardly construed as the paradigmatic ideal theory of justice—satisfies conditions (1)-(3), however, it also satisfies condition (9) and (10). Rawlsian ideal principles of justice are fact-sensitive because the contracting parties’ selection of such principles is informed by a comprehensive understanding of social science and economic theory.

Essentially, these different characteristic features of ideal and nonideal theory can either come apart from one another or cut against one another. Therefore, these characteristic features cannot be combined into a unified conception of “ideal” and “nonideal” theory, which neatly carves out the conceptual space.

---


9 John Rawls, A Theory of Justice: Revised Edition (Cambridge Mass.: Belknap Press, 1999), 119. (Of course, some might argue that it is more appropriate to label Rawls’s theory of justice using (4) rather than (9): Rawls states that the contracting parties should have a comprehensive knowledge of social science; however, perhaps, it is not accurate to say that the theory of justice that Rawls actually defends is really informed by rigorous social science.)
1.2.2 Deeper problems with the literature

In fact, it is not merely the case—given how philosophers have used the terms—that the demarcation between ideal and nonideal theory is somewhat messy and indeterminate. There are even deeper problems with some of the literature on this topic.

In the course of his attempt to “refute” ideal theory, Colin Farrelly argues that the following factors should be used to determine whether a theory is classified as ideal: inappropriate fact insensitivity, the incorporation of abstract hypotheticals, a cost blind approach, ignorance of constraints, idealization, and ignorance of what is feasible.\(^{10}\)

There are two distinct problems with Farrelly’s approach; each of these problems suggests general lessons. First, in defining ideal theory using so many distinct criteria Farrelly makes the distinction between ideal and nonideal theory incredibly indeterminate. How, for instance, should we classify a theory that satisfies some but not all of these criteria? As I noted above, it is true that different philosophers have proposed different conceptions of ideal and nonideal theory. However, an adequate conception must make things clearer rather than exasperating the present indeterminacy. Second, in partially construing ideal theory in terms of inappropriate fact insensitivity, Farrelly builds in a pejorative evaluative judgment about the value of ideal theory into his conception of ideal theory.\(^{11}\) This is inadequate: a conception of the subject matters of ideal and nonideal theory should be neutral, so that philosophers are in an adequate position to assess the respective virtues and vices of each type of theorizing. (Farrelly’s approach, in particular, makes his substantive conclusion about ideal theory—that it is valueless—rather trivial.)

The fact that different philosophers have defined ideal and nonideal theory in different senses has inevitably resulted in there being a number of merely verbal disputes in the literature. Jacob Levy’s recent work is perhaps the most dramatic illustration of this fact. He writes:


\(^{11}\) As Zofia Stemplowska notes, Farrelly commits himself to the evaluative claim that such fact insensitivity is “inappropriate” because he acknowledges “…that some fact insensitivity is necessary in order to avoid normatively privileging the status quo.” (Zofia Stemplowska, “What’s Ideal About Ideal Theory?,” *Social Theory and Practice* 34 (2008): 319-340, p. 320, fn. 5.)
The idea of normative political theory that is ideal in some absolute sense is a conceptual mistake, the equivalent of taking the simplifying models of introductory physics ("frictionless movement in a vacuum") and trying to develop an ideal theory of aerodynamics. Like aerodynamics, political life is about friction; no friction, no politics, no politics or justice.\(^{12}\)

Consequently, Levy concludes, all political theory is nonideal. However, Levy just stipulatively defines “ideal” theory in a particular sense and then uses this definition to argue that it is inappropriate for any philosopher to classify their theorizing as ideal in any sense. However, his argument (if successful) just shows that there is no ideal theory in the sense that he has stipulatively defined the term. It is irrelevant for the work of other philosophers (e.g., Rawls) who define ideal theory differently.

At most the dispute between philosophers who choose to label their work as “ideal” and philosophers such as Levy concerns the question of how it is appropriate to label types of theorizing. Levy clearly thinks, using analogies from science, that the label is misleading and inappropriate. However, nothing of substantive philosophical important hinges on how types of theorizing are labeled: in the context of political philosophy, ideal theory and nonideal theory are just terms of art. What is salient is whether such terms are a helpful way of illuminating interesting questions, not whether they diverge from different uses in other theoretical fields or everyday life.

1.2.3 Two strategies to overcome these problems

As David Schmidtz has perceptively observed, the contrast between ideal and nonideal theory has proved to be remarkably elusive.\(^{13}\) Consequently, it is not surprising that an increasingly number of philosophers have become discontent with the array of competing conceptions of ideal and nonideal theory scattered throughout the literature. In response, two different strategies have been proposed; both aim to carve out the conceptual space


more cleanly and thereby enable more fruitful theorizing about ideal and nonideal theory. First, philosophers such as Laura Valentini have argued that the different debates that have standardly been lumped under the single headings of “ideal vs. nonideal theory” should be delineated. This delineation will ensure that philosophers “do not talk past one another” and are able to carefully examine each independent component of the debate separately.14 Second, philosophers such as Gerald Gaus have tried to present a definitive conception of how we ought to use the terms ideal and nonideal theory. Such a conception is designed to latch onto the essential features of the debate better than the (somewhat sloppy) ways in which philosophers have standardly employed the terms.15

1.3 My approach

The approach I take lies somewhere in between these two strategies. I propose a way of defining the terms ideal and nonideal theory, and argue that my approach has certain theoretical virtues and avoids obvious pitfalls. However, I do not present this conception as definitive—I recognize that different conceptions of ideal and nonideal theory could also be perfectly reasonable.

1.3.1 Definition of ideal and nonideal theory

In the context of political justice, I define the subject matter of ideal and nonideal theory/justice exclusively in terms of (1) and (6): when we theorize about justice there are two different types of question that we can ask. First, how should a perfectly just institution (e.g., the basic structure of a domestic society) be arranged? Second, what should we do in

14 Laura Valentini, “Ideal vs. Non-Ideal theory: A Conceptual Map,” Philosophy Compass 7 (2012): 654-664, pp. 654-5. (She argues that the different debates that should be delineated are the following: (i) full vs. partial compliance (ii) utopian vs. realistic theory, and (iii) end state vs. transitional theory.) (Cf. Alan Hamlin & Zofia Stemplowska, “Theory, Ideal Theory and the Theory of Ideal,” Political Studies Review 10 (2012): 48-62, pp. 48-9.) They argue that the different debates that should be delineated are the following: (i) full vs. partial compliance (ii) idealization v. abstraction (iii) fact insensitivity v. insensitivity, and (iv) perfect justice v. local improvements.)

conditions that fail to fully realize perfect justice? If we try to answer the first question we are in the business of constructing an ideal theory of justice; if we try to answer the second we propose a nonideal theory of justice.\footnote{My explication of the distinction between ideal and nonideal theory follows one way in which commentators have standardly interpreted Rawls. (See Simmons, “Ideal and Nonideal Theory,” 7; Phillips, “Reflections on the Transition from Ideal to Non-ideal Theory;” 551.) It is also the same as one of a number of possible ways in which Valentini argues that we can understand the distinction: “ideal theory” (as I have defined the term) is what she refers to as “end-state theorizing”; “nonideal theory” (as I have defined the term) is what she refers to as “transitional theory.” (Valentini, “Ideal vs. Non-Ideal theory: A Conceptual Map,” 654-5.) Amartya Sen refers to “ideal theory” (as I have defined the term) as a “transcendental approach.” (Sen, “What Do We Want From a Theory of Justice?,” 216.)}

This way of construing the subject matter of ideal and nonideal theory/justice has some obvious virtues. It is clear and clean. It also does not build in any evaluative judgments about the value of ideal and nonideal theory and, consequently, puts us in a good position to assess the respective merits of both types of theorizing.

1.3.2 Clarification of related concepts

There is a striking asymmetry: every (reasonable) philosopher who defends the value of ideal theory also simultaneously maintains that nonideal theory is extremely important.\footnote{Rawls, for instance—the most famous exponent of an ideal theory of justice—continually recognizes the practical and theoretical importance of the pressing problems of nonideal theory. (See Rawls, \textit{A Theory of Justice}, 8, 343.)}

In contrast, a growing number of philosophers think that we should dispense with ideal theory entirely and just do nonideal theory.\footnote{See, for instance: Anderson, “Towards a Non-Ideal, Relational Methodology for Political Philosophy;” Farrelly, “Justice in Ideal Theory: A Refutation.”} In this dissertation, the term “nonideal methodology” refers to a methodology that exclusively supports nonideal theory.

I use the term “ideal conditions” to refer to states of affairs that realize ideal justice, “nonideal conditions” to refer to states of affairs that fail to realize ideal justice.
1.3.3 An alternative conception: strict v. partial compliance

Rawls sometimes characterizes the distinction between ideal and nonideal theory in terms of strict and partial compliance.\(^{19}\) Given that I am going to use a Rawlsian contractualist framework to generate a nonideal theory of justice, it might seem natural for me to follow Rawls and characterize the subject matter of ideal and nonideal theory in terms of strict and partial compliance. I argue, however, that there are good reasons not to take such an approach.

What Rawls means by strict and partial compliance is not entirely clear. One way of interpreting the distinction is in terms of the different methods used to construct ideal and nonideal theory: ideal theory is built on the idealizing assumption that actual people will strictly comply with the principles selected; nonideal theory dispenses with this idealizing assumption.\(^{20}\)

This way of distinguishing between ideal and nonideal theory has the virtue of being clear. However, the idealizing assumption of strict compliance that Rawls uses to derive ideal principles of justice is controversial and has been extensively criticized.\(^{21}\) Furthermore, it is clearly not necessary to adopt this idealizing assumption in order to offer a conception of how a perfectly just society should be structured. (Non-Rawlsian conceptions of perfect justice (e.g., feminist, utilitarian etc.,) are not derived against the backdrop of such an assumption). I submit that a plausible desideratum for a conception of the subject matter of ideal and nonideal theory is that, other things being equal, it should not make this conception depend on a controversial feature of a particular approach that a philosopher takes. If the subject matter of ideal and nonideal theory were picked out in terms of a controversial methodological assumption that Rawls makes, it would tie the conception of such theorizing too closely to a controversial feature of Rawls’s approach. Consequently, it is sensible for

---


even Rawlsian philosophers to construe the subject matter of ideal theory in terms of the comparatively neutral “conception of perfect justice.” This, in turn, will then put them in a position to argue that the best way of developing such a theory is to use the idealizing assumption of strict compliance.

A different way of interpreting the distinction is not in terms of the method used to derive ideal and nonideal theory but rather in terms of what would be required to realize or fail to realize ideal theory: ideal theory is realized if and only if all agents strictly comply with ideal theory. Nonideal conditions obtain if and only if at least a subset of agents do not comply; nonideal theory is an account of what ought to be done in such conditions.22

However, the problem with this approach is that all agents’ strict compliance with ideal principles of justice is neither necessary nor sufficient for a society to be ideally just in a plausible and clearly demarcated sense. This point is clearest within a broadly Rawlsian framework. As Thomas Pogge notes, one of Rawls enduring contributions to political philosophy is his re-conception of justice as “the first virtue of social institutions (i.e., the social system’s practices or “rules of the game” in its basic structure), rather than in terms of actions and consequences of actions performed by individual and collective agents.23 For Rawls—given that the primary subject of justice is the basic structure—the strict compliance of all actual agents is not sufficient to insure that the rules of the game in its basic structure are ideally just; relatedly—a point that I will expand on in Chapter 2—the realization of ideal justice is compatible with the noncompliance of at least some agents.

It is important to clarify that this conclusion does not depend on distinctive or controversial features of Rawls’s position. Even if Rawls’s claim that justice is the first virtue of social institutions is rejected, it is still highly plausible to think that the question of whether an institution is ideally just cannot be fully explicated in terms of facts about the compliance of actual agents. For instance, the question of whether fundamental constitutional laws are ideally just is irreducible to facts about the compliance of actual agents.

It is true that the degree to which agents are noncompliant is greater in nonideal conditions than in ideal conditions. However, given that this is merely a difference in degree, the

---

distinction is not as clean as Rawls’s labels of “strict” and “partial” compliance misleadingly suggest. It would seem that the only way of retaining the concepts of strict and partial compliance is to argue that in order to realize ideal justice, the basic structure of society—rather than actual agents—must strictly comply with the ideal principles of justice. However, this stretches the concept of compliance into the realm of the metaphorical: institutional rules are personified as agents capable of strictly or partially complying with certain principles. Given that the concept of compliance has to be rendered in this metaphorical way I think that it should be avoided because it is inelegant and potentially invites obscurity. This seems particularly the case once it is appreciated that in this metaphorical way it is, essentially, equivalent to the way I use the terms. For instance, a conception of how a perfectly just society should be structured is the same as a basic structure that strictly complies with ideal principles of justice.24

1.3.4 Final comment about terminology

In the emerging literature on nonideal theory, the question of how to characterize the distinction between ideal and nonideal theory is clearly deeply contested. This perhaps points to a deeper point: philosophers are experimenting with the conceptual space, in order to determine which way of characterizing the distinction allows them to explore the most substantive and interesting philosophical questions.

Such experimentation seems reasonable because the literature on ideal and nonideal theory is emerging. However, some of the literature seems haunted by the impulse to produce a definitive answer to the question “what is ideal and nonideal theory?” This is unfortunate, because the idea of some ultimate best way of defining terms of art is a chimera. At best definitions of such terms can be clear, illuminating and sensible; at worst unclear, confusing and eccentric. I think that my approach has the virtue of being clear. Whether it illuminates an interesting and important question will at least partially depend on the success

---

24 I think that the distinction between partial and strict compliance is a better way of framing certain moral questions—concerning the obligations of individual agents—than political questions concerning the demands of justice. For instance, the moral question of whether the non-compliance of a subset of agents should impose additional moral demands on compliant agents. (For a discussion of this interesting question, see: Liam B. Murphy, *Moral Demands in Nonideal Theory* (Oxford: Oxford University Press, 2000).)
of my arguments in the rest of this dissertation. It seems highly plausible to me that there are alternative, coherent ways of characterizing the distinction between ideal and nonideal theory that would facilitate the exploration of other interesting questions.  

1.4 Rawls’s sketch of nonideal theory

The contemporary distinction between ideal and nonideal theory has its origins in Rawls’s seminal work *A Theory of Justice*. In this and subsequent works, Rawls’s primary focus was on developing an ideal theory of justice. However, he also briefly explored some nonideal topics such as civil disobedience, and defended various conceptions of the content and structure of nonideal theory. In what follows I survey his conception of nonideal theory, and uncover some limitations and problems with his position as it stands.

1.4.1 The content and structure of Rawls’s nonideal theory

A. John Simmons offers the most comprehensive reconstruction of Rawls’s nonideal theory in the present literature. Drawing on *The Law of Peoples*, Simmons argues that Rawls’s nonideal theory has the following content and structure:

The specific “policies and courses of action” it mandates must be (i) “morally permissible,” (ii) “politically possible,” (iii) “likely to be effective” in moving society toward the ideal of perfect justice.

Condition (iii) captures the fact that Rawls’s nonideal theory is transitional: in nonideal conditions we should instigate courses of action that facilitate a transition towards ideal

---

25 One of the most interesting—and to my knowledge, entirely neglected—alternative ways of characterizing the distinction is suggested by Plato, in the following passage from the from *The Laws*: ‘This is particularly true of our legislation, where it looks as if we have a choice: either we can examine ideal laws, if we want to, or again, if we feel like it, we can look at the minimum standard we are prepared to put up with.’ (Plato, *The Laws*, trans. Trevor J. Saunders (London: Penguin Classics, 1975), 858a.) This suggests a conception of nonideal theory in which we try to articulate minimum, necessary conditions of permissibility.


27 Simmons, “Ideal and Nonideal theory,” 18.
justice. Conditions (i) and (ii) impose restrictions on such a transition. Condition (ii) is ambiguous. It could be interpreted as a minimal necessary condition, which simply rules out transitional paths that are impossible. Alternatively, it could be understood as a weightier condition that requires candidate transitional path to achieve a specified threshold of political feasibility. Condition (i) is the normative component of Rawls’s nonideal theory: there are constrains about what courses of action are acceptable that exceed mere political possibility.

1.4.2 Limitations of the permissibility condition

Of the three conditions, the permissibility condition, (i), is the most philosophically interesting because it is normative. Conditions (ii) and (iii) are empirical considerations that mainly require judgments of a social-scientific sort.28

As Simmons notes, Rawls does not expand on the permissibility condition.29 Therefore, the obvious limitation of this condition is that Rawls provides no real guidance as to which transitional paths are in fact permissible. Perhaps, he intends for legislators and citizens to rely on their intuition to rule out certain possible transitional paths. This certainly can be done in some extreme cases: for instance, it seems obvious that we should deem impermissible despotic rule by a dictator, even if this would be an effective long-term means of bringing about ideal justice. It is not, however, always so easy to determine the permissibility of a possible transitional path. There are a range of candidate policies and courses of action that give rise to conflicting reasonable intuitions. For instance, as I will argue in chapter 5, non-meritocratic affirmative action is a difficult case for nonideal theorists. On one hand, such a policy can seem appealing if it is an effective means of overcoming legacies of injustice (e.g., racial injustice). On the other hand, the following prima facie plausible objection can also be raised against such a policy: it imposes unfair burdens on individuals who are not members of the underrepresented groups that such policies benefit. In order to determine whether such a policy is permissible we require a more normatively determinate conception of permissibility. Rawls’s conception of the permissibility condition—at least as it stands—faces what I call the problem of “under-theorization:” it fails to provide a sufficiently rich and clear theoretical account.

---

28 Following Ibid. 19.
29 Ibid. 20.
A diagnosis as to why there are conflicting reasonable intuitions—and why it is therefore necessary to develop a more normatively determinate conception of permissibility—can be presented. Rawls does not tell us how conditions (i)-(iii) are to be weighted and, therefore, ‘…we are left to simply weigh intuitively the various desiderata as best we can.’ Simmons suggests that there is a plausible case to be made for the view that we cannot reasonably expect anything less sloppy and perhaps, ‘…it is enough to be able simply to articulate the various considerations—or “moral vectors”—that must be taken seriously in any reasonable defense of policy in nonideal circumstances.’

There is at least some plausibility to Simmons’s suggestion. For instance, it would probably be unreasonable to expect any nonideal theorist to develop a precise algorithm that explained how the social scientific judgments required by (ii) and (iii) are to be combined with the normative judgments required by (i) in order to generate specific policies.

However, even granting this, many conflicting reasonable intuitions can, I believe, be explained by the potential tension between conditions (i) and (iii). Condition (iii) requires us to effectively transition towards the realization of ideal justice. However, some possible transitional paths would raise the concern that people are treated unjustly, as a mere means to realize ideal justice for different people in the future. Such an instrumental use of agents stands in an obvious tension with one of the foundational aims of Rawls’s whole project: to develop a conception of justice that is sensitive to the separateness and inviolability of personhood. The permissibility condition, (i), is of course designed to counteract this concern. However, introducing strong constraints of permissibility—perhaps the constraints of permissibility that are prima facie judged necessary to avoid treating people in a problematic instrumental way—could give rise to different concerns. This is because the stronger the constraints of permissibility are, the harder it will be to realize the ideal; if the constraints are

30 Ibid. 19.
31 Ibid. 20.
32 As Rawls famously explains: ‘Each person possess an inviolability grounded on justice that even the welfare of society as a whole cannot override.’ (Rawls, A Theory of Justice, 3.) If this statement precludes using people at a particular time for the welfare of the majority at the same time, surely (it is also natural to argue) it precludes using either a subset of the population or the whole of the population at a particular time for the good of different people in the future.
sufficiently strong it might be almost impossible, in practice, to even partially transition towards the ideal.

Essentially, conditions (i) and (iii) elegantly capture two central considerations that should inform policy decisions in nonideal conditions: the constraint of treating people justly in such conditions, and the aim of transitioning to a more just social world. These considerations—whilst independently well motivated—will in practice often be in tension.

Given that the permissibility condition—and its relationship, in particular, with considerations of transitional effectiveness—are so under-theorized, Rawls’s conception of nonideal theory is silent on a number of central normative questions. For instance, to what extent, if any, can we violate considerations that animate the selection of ideal principles of justice (e.g., the separateness and inviolability of personhood) in order to effectively transition towards ideal justice in the long-term?\(^{33}\) We clearly need to know more about the content and strength of the permissibility condition in order to tackle this interesting and difficult question.

In summary, Rawls’s claim that nonideal policies and courses of action must satisfy the condition of permissibility is highly plausible. However, this merely seems like an

\(^{33}\) I think that we should be cautious of assuming that any prima facie normative tension between Rawls’s ideal and nonideal theory is ultimately problematic. After all, nonideal conditions are different from ideal conditions. Someone who defends an ideal although argues that the appropriate transition to that ideal can use means that would be prohibited after the ideal had been achieved does not contradict herself. An interesting illustration of this point is Lenin’s (in)famous discussion of disarmament: ‘Disarmament is the ideal of socialism. There will be no wars in socialist society; consequently, disarmament will be achieved. But whoever expects that socialism will be achieved without a social revolution and the dictatorship of the proletariat is not a socialist. Dictatorship is state power based directly on violence. And in the twentieth century—as in the age of civilisation generally—violence means neither a fist nor a club, but troops. To put “disarmament” in the program is tantamount to making the general declaration: We are opposed to the use of arms. There is as little Marxism in this as there would be if we were to say: We are opposed to violence!’ (Vladimir Ilyich Lenin, “The “Disarmament” Slogan,” in Lenin: Collected Works, Vol. 23, trans. M. S. Levin et al. (Moscow: Progress Publishers, 1964), 94.) It is at least prima facie tempting to think that there is a very problematic tension between Lenin’s ideal and nonideal theory: there seems to be at the very minimum a disquieting irony about achieving disarmament through violence. However, Lenin could protest that he is not contradicting himself: the adoption of a goal is one question; the permissible path of transition to that goal is a separate question. As there is at least no obvious contradiction, careful philosophical reflection is required about how (if at all) the adoption of a normative goal constrains the permissible path of transition to that goal.
uncontroversial starting point for any plausible nonideal theory.\textsuperscript{34} Indeed, it is tempting to suggest that until more details are filled in Rawls—and Rawlsians more generally—are not really presenting a substantive nonideal theory of justice.

1.4.3 Additional problems with the permissibility condition given Rawls’s methodology

It is not merely the case that Rawls’s conception of the permissibility condition faces the problem of under-theorization. Given his distinctive methodological commitments he faces at least two additional problems.

\textit{Priority rules}

In the context of ideal theory, Rawls adopts “priority rules”, which specify how different principles and considerations are to be weighted against one another.\textsuperscript{35} In fact he even goes so far as to state that:

The assignment of weights is an essential and not a minor part of a conception of justice. If we cannot explain how these weights are to be determined by reasonable ethical criteria, the means of rational discussion have come to an end.\textsuperscript{36}

\textsuperscript{34} The only way to challenge such a starting point is to argue that there are, at least in some instances, unavoidable dilemmas such that certain courses of action both should and should not be pursued. (For a good defense of such a claim, see: Michael Walzer, “Political Action: The Problem of Dirty Hands,” \textit{Philosophy & Public Affairs} 2 (1973): 160-180.)

\textsuperscript{35} Rawls, \textit{A Theory of Justice}, 37.

\textsuperscript{36} Ibid. Rawls’s innovation and defense of priority rules plays a crucial part in his attempt to provide a systematic and plausible alternative to Utilitarianism. Intuitionistic moral theories set out an irreducible family of first principles that have to be weighted against one another intuitively. Rawls concluded that such theories are vulnerable to the objection, voiced by Utilitarians such as Mill and Sidgwick, that at some point we must have a single principle to straighten out and systematize our judgments and that the principle of utility maximization was the only plausible principle that could perform this requisite function. (Ibid. 36) Rawls, like Intuitionists, believed that there are multiple moral principles. However, he argued that he could consistently retain this commitment to multiple moral principles and overcome the Utilitarian challenge by introducing
As I noted above, Rawls does not flesh out the normative content of the permissibility condition. Consequently, he is not in a position to specify how the different normative considerations encompassed by this condition are to be weighted against one another. It is, however, important to emphasize—assuming that there are multiple normative considerations—that his conception of the permissibility condition is problematic not merely because it is not rich enough to offer adequate normative guidance but also because it is not structured by priority rules.

The same set of reasons that motivate the adoption of priority rules also apply in the context of nonideal theory: priority rule structure the different demands of justice into a clear set of public standards; such standards define, in particular, the political responsibilities of citizens. Indeed, if anything, the function that such rules perform of clearly structuring political responsibilities might well be more important in nonideal conditions, in which such responsibilities are likely to be both more contested and onerous.

Furthermore, Rawls maintains that priority rules are an essential part of an ideal theory of justice. However, in the context of nonideal theory he does not supply priority rules. Consequently, there seems to be a problematic methodological tension between Rawls’s conception of ideal and nonideal theory. It is after all difficult to see what principled reason could be supplied for why the methodological requirement of priority rules is an essential feature of ideal theory, although can be completely abandoned in the context of nonideal theory. This methodological tension puts Rawls in a dialectically weak position against Intuitionists—who maintain that different normative considerations need not be ordered using priority rules. Such Intuitionists can justifiably complain that Rawls should not object to Intuitionism in the context of ideal theory if Rawls himself slides into adopting an Intuitionist methodology in the context of nonideal theory.

priority rules. He famously argued that the ideal principles of justice must be put in “serial or lexical order.” ‘This is an order which requires us to satisfy the first principle in the ordering before we can move on to the second…A serial ordering avoids, then, having to balance principles at all.’ (Ibid. 38)


As I concluded above, Rawls’s conception of the permissibility condition is so under-theorized that there is a normative lacuna at the heart of his nonideal theory. Perhaps, it could be suggested, this is ultimately unproblematic: importing generic considerations of permissibility from broader moral theorizing can simply fill out this lacuna.

This suggestion seems plausible. However, given another one of Rawls’s central methodological commitments it is not a suggestion that a Rawlsian can embrace. Rawls himself, cautioned in *A Theory of Justice* that there is no *a priori* reason for thinking that ‘principles [of justice] satisfactory for the basic structure [of society] hold for all cases.’

Later in his philosophical career, when he wrote the paper “Justice as Fairness: Political not Metaphysical,” Rawls thought that he became clearer about the distinction between moral and political philosophy. However, whatever we may think about the plausibility of his later political turn—in which political philosophy is conceived as freestanding and independent from comprehensive moral doctrines—even in *A Theory of Justice* he makes the plausible claim that the normative content of principles of justice need to be sensitive to facts about the particular institutional structure that they are designed to regulate.

Therefore, Rawls’s failure to fill out the normative content of the permissibility condition is ultimately problematic. This is because it cannot be filled out by importing generic considerations of moral permissibility. Rather, there is a need to determine—with respect to the specific institutional context of the basic structure of society—what considerations should constrain our pursuit of ideal justice.

---

1.4.4 Problems with the relationship between Rawls’s ideal and nonideal theories

A crucial component of Rawls’s nonideal theory is his conception of its relationship with ideal theory. Throughout his career, Rawls focused on ideal theory and thought that he did so with good reason. He argued that nonideal theory was secondary and dependent on ideal theory: an ideal theory, he believed, provides ‘…the only basis for the systematic grasp of these more pressing problems [of nonideal theory].’ In the following, important passage Rawls elaborates on the relationship between ideal and nonideal theory:

By putting these principles [of ideal justice] in lexical order, the parties are choosing a conception of justice suitable for favorable conditions and assuming that a just society can in due course be achieved. Arranged in this order, the principles define then a perfectly just scheme; they belong to ideal theory and set up an aim to guide the course of social reform. But even granting the soundness of these principles for this purpose, we must still ask how well they apply to institutions under less than favorable conditions, and whether they provide any guidance for instances of injustice. The principles and their lexical order were not acknowledged with these situations in mind and so it is possible that they no longer hold.

Rawls implies that—at least under reasonably favorable nonideal conditions—nonideal theory depends on ideal theory: in such conditions we should transition to the full realization of ideal justice, and this transition should be orientated by ideal theory. He acknowledges, however, that the applicability of ideal theory to nonideal conditions may be limited in certain respects. For instance, there is no guarantee that—particularly in a lexically ordered form—ideal principles will be a reliable guide in less than favorable conditions.

This acknowledgement seems reasonable. However, there is a deeper skeptical challenge, here, that Rawls fails to adequately appreciate. This is because just as there is no a

---

43 Ibid. 215-216.
44 I will here put to one side the concern that Rawls’s uses of the terms “reasonably favorable conditions” and “less than favorable conditions” are incredibly vague.
priori guarantee that ideal theory applies to nonideal conditions under less than favorable conditions in certain respects, there seems to be no a priori guarantee that it applies to nonideal conditions under reasonably favorable conditions in any respect.

It may seem that ideal theory must directly apply to nonideal conditions under reasonably favorable conditions: given that ideal theory sets out a conception of how the world should be, it must be something that we should try to realize under nonideal conditions—at least if we are in reasonably favorable conditions in which this can be done. Caution, however, is required before we unreflectively adopt this position. We can consistently adopt the claim that the world should be a certain way, whilst arguing that we should not try to realize it in nonideal conditions that are reasonably favorable because a different set of normative considerations arise within even reasonably favorable nonideal conditions. In particular, we might argue that we should simply try to remedy instances of actual injustice without (necessarily) bringing about ideal, future justice. Alternatively, in nonideal conditions ideal theory may not perform the function of guidance: perhaps people can muddle along improving their social world, without needing to draw guidance from ideal theory.

To be sure Rawls goes on to explain:

Viewing the theory of justice as a whole, the ideal part presents a conception of a just society that we are to achieve if we can. Existing institutions are to be judged in the light of this conception and held to be unjust to the extent that they depart from it without sufficient reason. The lexical ranking of the principles specifies which elements of the ideal are relatively more urgent, and the priority rules this ordering suggests are to be applied to nonideal cases as well.  

However, Rawls is simply asserting here that considerations of ideal justice directly apply to nonideal conditions at least under reasonably favorable conditions. He has not, however, presented a justification for this claim and has not, therefore, overcome the above skeptical challenge.

---

45 Rawls, A Theory of Justice, 216.
Even putting aside this general skeptical challenge, further obscurity is introduced by the fact that Rawls does not clarify the precise way, or ways, in which nonideal theory is supposed to depend on ideal theory, and in fact advances different conceptions of the relationship.

Simmons argues that we should see Rawlsian nonideal theory as strongly transitional and seeking to realize an integrated rather than piecemeal ideal of perfect justice. By “integrated” Simmons explains that he means ‘…good policies are good not relative to the elimination of any particular, targeted injustices, but only relative to the integrated goal of eliminating all injustice.’ Simmons’s interpretation has considerable exegetical support, including the passages I quote above. Christine Korsgaard advances a different influential conception of the relationship: Rawls’s nonideal theory must be constrained by features of his ideal theory. Note that Korsgaard’s interpretation is compatible with that of Simmons: it maybe the case that we should transition toward ideal justice, and that this transition should be constrained by features of ideal theory.

However, things are complicated by the fact that Rawls, himself, sometimes conceives the relation between ideal and nonideal theory completely differently. In the context of his discussion of economic ideals such as “a perfectly competitive economy”, which Rawls acknowledges are not actually realized in our nonideal world, he argues that such economic ideals help us ‘…to decide upon the best way to approximate the ideal.' The concept of “approximating an ideal” seems to mean that at any particular time we should try to implement institutions—in our actual nonideal world—that partially (perhaps as extensively as possible?) resemble ideals. This concept is, therefore, distinct from Simmons’s claim that we should transition to an integrated ideal, which means that at any particular time we should not try to approximate an ideal of justice as directly as possible; rather we should transition—even if at a particular time this requires taking a step away from the ideal—in the most effective and permissible way towards an integrated ideal of justice.

In summary, Rawls does not supply a clear conception or justification of the relationship between ideal and nonideal theory; in the secondary literature this relationship is the

---

46 Simmons, “Ideal and Nonideal Theory”, 22. (Cf. Rawls, A Theory of Justice, 216.)
subject of ongoing dispute. In 1985, Michael Philips observed: ‘…neither Rawls nor any of his followers have clarified the nature of the derivation [i.e., the dependence of nonideal theory on ideal theory] in any depth or detail.’\textsuperscript{49} I think it is fair to say that since then some progress has been made, although it still remains a crucial albeit surprisingly obscure and undertheorized component of Rawls’s work.

1.5 Objections to Rawls’s position

Given that Rawls’s conception of nonideal theory and its relationship with ideal theory are so under-theorized and inconsistent, it is not surprising that it has been subjected to such extensive criticism. Indeed, an ever-growing number of philosophers have concluded that because of seemingly intractable problems with this aspect of Rawls’s theory, we should abandon the whole paradigm of Rawlsian justice.

1.5.1 Amartya Sen’s objection

Amartya Sen—who is perhaps Rawls’s most prominent critic in this regard—argues that in order to determine what should be done in nonideal conditions, it is not necessary to antecedently determine an ideal theory of justice.\textsuperscript{50} Therefore, rather than getting heated up about what the content of ideal justice requires, political philosophers should adopt a nonideal methodology and, thereby, exclusively focus on crafting solutions to the pressing questions of nonideal theory that actually confront them.\textsuperscript{51}

\textsuperscript{49} Philipps, “Reflections on the Transition from Ideal to non-Ideal Theory,” 551.

\textsuperscript{50} Sen, “What Do We Want From a Theory of Justice?,” Sen, The Idea of Justice. (Cf. Anderson, The Imperative of Integration, 3-7.) Sen also argues that an ideal theory of justice is not sufficient to determine what ought to be done in nonideal conditions. However, like Pablo Gilabert, I think that this challenge should be sidelined. This is because no reasonable defender of ideal theory thinks that ideal theory all on its own is sufficient to establish what institutional arrangements ought to be adopted in nonideal conditions; whereas many philosophers, such as Rawls, and Simmons argue that ideal theory is necessary to make judgments of this sort. (Pablo Gilabert, “Comparative Assessments of Justice, Political Feasibility, and Ideal Theory,” Ethical Theory and Moral Practice 15 (2012): 39-56, p. 44.)

\textsuperscript{51} Cf. Anderson The Imperative of Integration, 3-7.
Two quite different considerations seem to motivate Sen’s objection. First, *prima facie* plausible counter-examples can be produced to the claim that nonideal theory necessarily presupposes ideal theory. For instance, we know that slavery is wrong prior to determining the normative content of ideal theory. Relatedly, philosophers who endorse different ideal theories (e.g., feminist, Lockean, and Rawlsian) can reach a consensus on certain topics in nonideal theory, such as the viability of Universal Basic Income, whilst disagreeing about the content of ideal theory.\(^{52}\)

Given that ideal theory does not appear to be strictly necessary, and Rawls’s conception of the relationship between ideal and nonideal theory is so messy and indeterminate, it is easy to feel sympathy for Sen’s position: Rawls’s theory of justice is baroque and architectonic, and its application to actual political practice in our nonideal world is obscure; consequently, perhaps it is sensible to turn away from this baroque system and begin afresh by tackling topics of nonideal theory independently from ideal theory.

The second consideration that motivates this objection is a deeper, existential complaint about the Rawlsian approach. Sen notes that the basic orientation of Rawlsian ideal theory is completely different from the types of concern that engage actual people who are trying to remedy instances of injustice in the real world.\(^{53}\) Indeed, Sen prefaces his book *The Idea of Justice* by noting that ‘[w]hat moves us, reasonably enough, is not the realization that the world falls short of being completely just—which few of us expect—but that there are clearly remediable injustices around us which we want to eliminate.’\(^{54}\)

If we are moved by these concerns, we might view Rawlsian ideal theory as an intellectually lethargic—even blameworthy—approach to theorizing about justice: why should we toy with abstract models of how the world should ideally be, like Nero fiddling while Rome burns,\(^{55}\) when we are confronted with so many abhorrent and rampant instances of actual injustice; surely our creative, intellectual energies should be focused on trying to remedy

---


\(^{53}\) Ibid. 218.


instances of actual injustice rather than on articulating picturesque ideals of how the world should be.\textsuperscript{56}

1.5 Two Rawlsian options and my plan moving forward

I will close by outlining two different ways of responding to the problems with Rawls’s position as it stands. I will motivate the approach that I adopt, and explain its value both for Rawls’s theory of justice and non-Rawlsian work on ideal and nonideal theory.

1.5.1 Two Rawlsian options

The first option is to retract, or remain neutral, about the claim that nonideal theory depends on ideal theory, and to argue that ideal theory has intrinsic value that is independent of whatever instrumental value it may offer, such as its ability to orientate actual policy decisions. This is the position that David Estlund occupies. According to Estlund, ideal theory is important and it is valuable therefore to come to understand it for that reason.\textsuperscript{57}

Gaus acknowledges that it is possible that ideal theory may have intrinsic value even if it is not instrumentally valuable. However, he derogatorily labels this a “mere dreaming

\textsuperscript{56} Although Swift seems to sympathize with the complaint that political philosophers have not been sufficiently preoccupied with the practical task of making the world less unjust, he expresses some perplexity when this type of objection is presented as the judgment that philosophers who construct ideal theories of justice that fail to offer guidance are somehow blame\textsuperscript{57}eworthy. He writes: ‘It is striking that we are less likely to criticize, violinists, say, than political philosophers, for failing to provide justice-promoting guidance, as if being interested in identifying truths about justice meant that one was more rather than less culpable for failing to tell us how to bring it about.’ (Ibid. 367). I find Swift’s suggestion provocative and interesting. However, I think that it could reasonably be objected that although political philosophers are not more blameworthy \textit{qua} human beings than violinists, they are blameworthy \textit{qua} theorists about justice, at least in a certain respect. This is because in their capacity \textit{qua} theorists, they arguably have an intellectual and social duty to direct their energies to questions of social importance and relevance.

“account” to highlight that this is a very minor acknowledgment to make; an acknowledgment, he believes, that is compatible with overcoming the orthodox ideal theory paradigm.\(^{58}\) Even if we do not share Gaus’s sentiments—indeed, even if like Estlund we celebrate the intrinsic value of ideal theory—this is an approach that “gives up” on a significant value of ideal theory: the indispensible instrumental function of orientation. Rawls, for instance, is always adamant that the value of ideal theory is at least primarily derivative. In *A Theory of Justice* he explains that “[i]f 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 nonideal part as well.”\(^{59}\) In fact, Rawls’s clearest statement of this idea is presented in the following passage from *Political Liberalism*:

> The work of abstraction, then, is not gratuitous: not abstraction for abstraction’s sake. Rather, it is a way of continuing public discussion when shared understandings of lesser generality have broken down.\(^{60}\)

I personally find Estund’s claim that ideal theory has intrinsic value compelling and plausible. However, I think it must be acknowledged if this is the only value that is has such value would be disappointing if we share Rawls’s grand ambitions for ideal theory.

The second option is to acknowledge that Rawls’s conception of nonideal theory is problematic as it stands, but to argue that these problems are not insurmountable because the central ambitions of his project can be successfully carried through. This is the option that I adopt in this dissertation. I use Rawls’s powerful contractualist framework to forge a *sui generis* nonideal theory of justice and clarify its interrelationship with his ideal theory of justice. In doing so, I clarify and defend the respects in which nonideal theory depends on ideal theory.

This is clearly a dauntingly ambitious project and the scope of my project is limited in a number of significant respects. In *A Theory of Justice* there is a separate ideal theory of justice for each of the following three things: the basic structure of a domestic society,

---


individual moral responsibilities, and international relations between nations. Each of these three ideal theories of justice has a corresponding nonideal theory. Within this dissertation I exclusively focus on the nonideal theory appropriate for the basic structure of a well-ordered, relatively affluent domestic society. My development of Rawlsian nonideal theory, even in this respect, is by no means exhaustive. In order to offer a complete account it would be necessary to extend the theory to particular topics such as punishment and to

---

62 Rawls's is explicitly clear that his theory of justice applies if and only if the circumstances of justice obtain. These circumstances of justice include objective factors, such as moderate scarcity, and subjective conditions, such as cooperation being psychologically possible. (See Rawls, *A Theory of Justice*, 109-12). One interesting ambiguity that I think we find in Rawls’s work is the scope of actual societies that satisfy the circumstances of justice. This ambiguity arises because concepts such as “moderate scarcity” and “cooperation being possible” are extremely vague. In recent work Colleen Murphy has argued that we cannot apply standards of justice that are appropriate to a relatively stable (although still clearly unjust) country such as the United States to transitional societies (i.e., societies that are attempting to make the difficult transition from oppressive non-democratic forms of rule to democracy, such as post-Apartheid South Africa) because the circumstances of justice that apply in such transitional contexts are different. (See Colleen Murphy, *The Conceptual Foundations of Transitional Justice* (Cambridge: Cambridge University Press, 2017).) For the purposes of this dissertation I will try to identify and defend principles of nonideal justice that are appropriate for a relatively stable democratic society such as the present day United States. I am neutral about the extent to which (if at all) this conception can be extended to different types of societies in general and the transitional societies in particular that Murphy discusses. Nevertheless, a useful byproduct of my analysis is that it will put us in a better position to assess Murphy’s claim that different considerations of justice apply to a transitional society (in her technical sense of the term) than the nonideal principles of justice that should govern a relatively stable democracy like the United State’s pursuit of ideal justice. This is because in order to draw any such contrast with confidence we must be clearer both about the standards of justice that apply to the present day United States and to transitional societies like present day South Africa.
63 This is not to say that I am unsympathetic to the familiar cosmopolitan critique of Rawls: the modeling of domestic societies as closed is problematic because considerations of domestic justice are inseparable from questions of international justice. However, given that the main aim of the dissertation is to develop a clear overview of the structure and content of nonideal theory and its relationship to ideal theory, it would overcomplicate things to also consider how demands of international justice intersect with those of domestic justice. I think that my theory could easily be adapted to accommodate the insights of the cosmopolitan critique of Rawls. (For a good presentation of the cosmopolitan critique of Rawls see: Thomas Pogge, “Cosmopolitanism and Sovereignty,” *Ethics* 103 (1992): 48-75; Thomas Pogge, “An Egalitarian Law of Peoples,” *Philosophy & Public Affairs* 23 (1994): 195-224.)
assess Rawls’s own extension of nonideal theory to the topic of civil disobedience. Furthermore, clearly much more must be said than I have space to say about how Rawlsian nonideal theory intersects with other things, in particular actual democratic practice and public reason. These topics and considerations all fall outside the scope of my dissertation; however, a satisfactory treatment of them all presuppose my project of developing a general conception of Rawlsian, nonideal theory.

1.5.2 The value of my project

The most obvious value of my project is that it develops an important, but under-theorized, aspect of Rawls’s work. A broadly Rawlsian position is still seen by many as an intellectually live option. However, because Rawls’s nonideal theory has never been adequately developed we currently only have a partial grasp of his theory of justice. I hope that developing a comprehensive nonideal theory of justice within a Rawlsian contractualist framework will put us in a better position to assess Rawls’s theory of justice as a whole.

I also hold out the aspiration that my project will have value for non-Rawlsian work on ideal and nonideal theory. Most obviously, this is because at least some of my arguments, for instance my argument for the agent-guiding function of justice in chapter 3, generalize outside a Rawlsian contractualist framework. More subtly, the ideal and nonideal theory debate is increasingly construed as a methodological debate. Zofia Stemplowska and Adams Swift note, for instance, that the contemporary debate about nonideal theory has spilled over from its Rawlsian origins into a dispute about the methods and aims of political philosophy in general. In one sense this clearly seems reasonable enough. Questions such as what level of idealization is appropriate in political philosophy, and whether nonideal theory presupposes ideal theory, clearly do not depend on the acceptance of substantive features of Rawls’s theory of justice. However, a dissatisfaction that I have with at least some of the literature on ideal and nonideal theory is that the attempt to articulate—in purely general, methodological terms—a conception of nonideal theory and its interconnection with ideal theory has yielded results that are quite meager and indeterminate. Even if we think that

---

there are *sui generis* questions about methodology they are probably best explored in tandem with first-order philosophical disputes.\textsuperscript{65}

I defend this claim further in chapter 5 (§5.8.2). I argue that the methodological question of what types of theorizing are instrumentally valuable should be settled by the following test: consider which types of theorizing are required to craft the most compelling philosophical accounts of first-order problems (e.g., When, if at all, is non-meritocratic affirmative action justice?). Essentially, the second-order methodological question should be determined by what an adequate treatment of the first-order problems requires. Given this, I will adopt the approach of developing a broadly Rawlsian nonideal theory of justice, which I argue has important implications for certain core first-order problems. I will then try to extrapolate some more general methodological conclusions out of this substantive theory, rather than trying to reach these methodological conclusions in independence from either the substantive normative content of nonideal theory, or an assessment of first-order problems.

1.6 Summary of chapter 1

In this chapter I have surveyed the—at times perilously messy and indeterminate—literature on ideal and nonideal theory. I have presented and defended conceptions of the subject matters of ideal and nonideal theory. I have shown that Rawls’s conception of nonideal theory and its relationship with ideal is underdeveloped and problematic. Finally, I have indicated how I will develop his theory, and the value that this has for both Rawlsian and non-Rawlsian work on ideal and nonideal theory.

\textsuperscript{65} This is the approach taken by other philosophers who have explored philosophical methodology in recent years. For instance, Ted Sider makes his case for ontological realism by exploring contemporary debates about mereology. See: Theodore Sider, “Ontological Realism” in *Metametaphysics* ed. David J. Chalmers, David Manley and Ryan Wasserman (Oxford: Oxford University Press, 2009), 384-421.
Chapter 2

NONIDEAL PRINCIPLES OF JUSTICE

2.1 Introduction

As I concluded in chapter 1, Rawls’s conception of nonideal theory and its relationship with ideal theory is both under-theorized, and thwarted by some deeper theoretical concerns. In this chapter I show how these problems can be overcome. I do so by using a Rawlsian contractualist framework to derive a set of nonideal principles of justice. The innovation of such principles is, I argue, the key to giving substantive normative content to the indeterminate “moral permissibility” condition that we inherit via Simmons from Rawls. Furthermore, such principles can be used to clarify the relationship between ideal and nonideal theory because they directly constitute this relationship.

This chapter has the following structure: first, I undertake the preliminary task of setting out a contractualist framework. Second, I derive and defend a set of nonideal principles of justice. Third, I clarify the relationship between ideal and nonideal theory. Fourth, and finally, I show how some of my analysis generalizes, even if particular assumptions that I adopt—such as the claim that Rawls’s ideal theory of justice is correct—are rejected.

2.2 Nonideal theory, contractualism, and reflective-equilibrium

In order to determine what should be done in nonideal conditions, it would be possible merely to appeal to considered intuitions or present principles that are intuited from a realist moral ontology. However, such methods clearly seem inappropriate for philosophers such as Rawls, who use a contractualist procedure to derive ideal principles of
justice. If it is impermissible to determine the content of ideal theory using such methods, then surely there is no reason why it is also unacceptable to determine the content of nonideal theory using such methods. I contend that the most plausible method for such philosophers is to use a contractualist framework to derive nonideal principles of justice.

Like Rawls, I think that a contractualist procedure should be embedded in a broader reflective-equilibrium methodology. The purpose of a contractualist procedure is, essentially, to model and extend our intuitions about justice, which we take as provisional fixed points. If we find a discrepancy between the outcome of the contractualist procedure that we construct and our considered intuitions about justice then we face a choice:

We can either modify the account of the initial situation [i.e., the contractualist procedure] or we can revise our existing judgments, for even the judgments we take provisionally as fixed points are liable to revision.

We thus zip back and forth, until we find a description of the contractualist procedure that yields principles which fully match our considered judgments about justice. A consequence of endorsing this reflective-equilibrium methodology is that my theorizing in this chapter is only provisional: if the nonideal principles that I generate do not harmonize with considered intuitions about nonideal theory, then one possible option is to tinker with the contractualist framework in order to generate a different set of nonideal principles of justice.

2.3 Conjecture about why Rawlsian nonideal theory has been neglected

Given the prominence of Rawls’s work within contemporary political philosophy and the recent surge of interest in nonideal theory, it is remarkable that so little work has been done trying to develop a nonideal theory of justice using a Rawlsian contractualist framework. This seems particularly perplexing given that many contemporary philosophers

---

have used Rawls’s theoretically powerful innovations of an “original position” and “veil of ignorance” to extend his theorizing to different domains. For instance, Susan Moller Okin has extended Rawls’s contractualist method to address considerations of sexual inequality within both the family and society; Charles Beitz has developed a contractualist theory of international relations and in doing so challenged a number of the substantive claims that Rawls, himself, makes in *The Law of Peoples.*

I would conjecture that one of the main reasons that Rawlsian nonideal theory has been so underexplored is that the most radical, contemporary exponents of nonideal theory, such as Charles Mills and Colin Farelly, have explicitly styled themselves in opposition to Rawlsian ideal theory. They view Rawlsian political philosophy as the historically dominant approach to justice that they want to overcome. I suggest that the standard way in which nonideal theory has been juxtaposed with Rawlsian ideal theory has resulted in the neglect of the possibility that a Rawlsian contractualist framework can be harnessed to forge a compelling nonideal theory of justice.

2.3 Survey of attempts to develop a nonideal theory of justice within a contractualist framework

Although the project of developing a nonideal theory of justice within a contractualist framework has been remarkably underexplored, Marcus Arvan and Charles Mills have (at least to some extent) undertaken such a project. Accordingly, I will begin by surveying their work. In doing so, I will explain why I take a different approach.

2.3.1 Marcus Arvan’s contractualist framework

Arvan attempts to derive a nonideal theory of justice using a Rawlsian contractualist framework. He argues that the nonideal contracting parties should be modeled in the following way:

---


5 It is worth noting that the ambitions of Arvan’s project are slightly different from that of my own. Arvan argues that: ‘[T]here are at least some nonideal, real-world conditions in which Rawlsian circumstances
‘[B]ecause nonideal conditions are less than fully just, and justice is a normative notion (viz. injustices are wrong), everyone under nonideal conditions must be modeled as having an equal obligation to prefer a fully just society and the elimination of any and every injustice (note that this fits nicely with Rawls’s own idea that everyone in nonideal conditions has ‘natural duties of justice’).\(^6\)

However, in response, ideal justice is also a normative notion. Yet, this does not require the ideal contracting parties to be modeled as having any antecedent obligations with respect to justice. Therefore, the fact that justice—in a nonideal sense—is a normative notion does not require the nonideal contracting parties to be modeled in such a way.

Furthermore, in building in an antecedent obligation to eliminate injustice into the contractualist framework Arvan deprives the contractualist procedure of its intended value: to provide an explanatory heuristic. One way in which the procedure should perform this heuristic function is to demonstrate that under nonideal conditions people ought to eliminate injustice. This function should be achieved—in the context of nonideal theory as in the context of ideal theory—by demonstrating that certain principles of justice would be freely adopted by individuals trying to advance their own individual ends behind the veil of ignorance.\(^7\) However, by antecedently assuming that parties have an obligation to eliminate of justice fail to obtain, and in which considerations of justice still intuitively apply.’ (Marcus Arvan, “First Steps Towards a Nonideal Theory of Justice,” *Ethics and Global Politics* 7 (2014): 95-117, p. 98.) Arvan gives the examples of failed states and Tim Mulgan’s examination of justice after an environmental catastrophe that has rendered normal circumstances of cooperation impossible, although it is still possible for people to cooperate to some extent. (See: Tim Mulgan, *Ethics for a Broken World: Imagining Philosophy After Catastrophe* (Acumen, 2011).) As I explained in chapter 1, my own development of nonideal theory presupposes economic conditions of relative affluence. However, even granting the fact that Arvan has different aims, I still think—for the reasons that I will present—that the contractualist framework that he uses is ill suited to achieve these aims.


\(^7\) Tim Scanlon explains this point well when he writes: ‘The resulting view, that the correct principles of justice are ones that would be agreed to by parties in the Original Position Rawls described, has the form of a constructivist view: facts about justice are facts about what principles would be arrived at through a process of a certain kind…[T]he account remains constructivist because the judgments in question are not judgments
injustice Arvan is unable to use the contractualist model to perform this heuristic function. The output “we ought to eliminate injustice under nonideal conditions” is trivial because it is an input.

More subtly, the point of a nonideal contractualist procedure is not merely to affirm a general commitment to eliminate injustice. Rather it should specify, in more detail, what ought to be done under nonideal conditions by determining the content and limits of the demands of justice. If the nonideal contracting parties are modeled, like the ideal contracting parties, as rational and disinterested, the output of the contractualist procedure has (at least the potential) to achieve this desired specificity. We can consider what principles of justice (and associated obligations) the contracting parties would freely choose to adopt by weighing the costs and benefits of candidate principles against one another. However, on Arvan’s model—in which the parties are modeled as having an obligation to eliminate each and every injustice—it is not possible to hone in on principles that give greater content to what should be done under nonideal conditions. Perhaps most significantly, given that the contracting parties are antecedently modeled as preferring the elimination of each and every injustice, they are ex hypothesi unable to consider the possibility that they would prefer not to eliminate an injustice, if the elimination of such injustice requires them to endure sufficiently high costs. I think that the inability for them to entertain such a possibility—right out of the contractualist gate—is disastrous. For, as I will argue below, restrictions of this sort are what the contracting parties would choose to adopt.

2.3.2 Charles Mills’s contractualist framework

Mills rejects Rawls’s ideal theory of justice; in fact, he disavows ideal theory in general because he judges it ideological. However, whilst rejecting central features of Rawls’s approach, he indicates that he wants to remain broadly within the social contract tradition. He argues that the social contract should be constructed as a covenant to end actual modes of social domination, such as racial injustice. In this regard he suggests that rather than

about what is just or unjust (or morally right or wrong) but rather judgments about what individuals who are seeking only to do as well for themselves as they can would have reason to choose under conditions of limited knowledge.’ (T. M. Scanlon, “The Appeal and Limits of Constructivism,” in Constructivism in Practical Philosophy, ed. James Lenman and Yonatan Shemmer (Oxford: Oxford University Press, 2012), 232.)
Rawls’s contractualist framework, Jean Jacques Rousseau’s “domination contract” in *Discourse on the Origins of Inequality* should be our model, ‘…since it directs our theoretical attention from the start to domination and exploitation as central to the sociopolitical order.’ Essentially, rather than adopting the Rawlsian approach of modeling an abstract conception of ideal justice—which in nonideal conditions we should transition towards—Mills suggests that we should use contractualism to focus on overcoming instances of injustice that afflict us in our actual nonideal world.

In contrast to Arvan’s work, there is not anything obviously problematic with Mills’s approach. Significantly, the nonideal principles of justice that I derive in this chapter are a possible elaboration of Mills’s envisioned social contract to end domination. For, they specify how we should overcome actual instances of justice, rather than how a perfectly just society should be structured. However, in contrast to Mills’s own elaboration of this approach in *The Racial Contract*, I do not think it is necessary to ditch either ideal theory or Rawls’s contractualist framework in order to carry out this project. Indeed, as I will argue below, I think that ideal theory is absolutely indispensable to carry off such a project.

2.4 My contractualist framework and preliminary distinctions

In what follows, I set out the contractualist framework I use to derive a set of nonideal principles of justice.

2.4.1 Symmetries between the ideal and nonideal original positions

For the reader acquainted with Rawls, the contractualist framework that I use to derive nonideal principles of justice will be largely familiar. This is because, in almost all respects, it is exactly the same as the one that Rawls uses to determine ideal principles of justice. The nonideal original position contractors are modeled as rational, in the sense that

---


they try to realize their ordered preferences as effectively as possible. Given this conception of rationality they are exclusively concerned with their own preferences; they are not altruistic. The parties are placed behind a veil of ignorance. This insures that the parties do not know particular facts about the social position that they will occupy in an actual society. They do not know their social status, relative intelligence, conception of the good, sex, or race, etc. This veil does not, however, preclude knowledge of the general facts about human society: the parties have a comprehensive grasp of social science and moral psychology, etc.

Essentially, the nonideal contractualist procedure—like the ideal contractualist procedure—is designed to set up a fair procedure. This insures that the principles selected using this procedure are just.

2.4.2 The intergenerational component

The core aspects of the contractualist framework are familiar. However, how the framework models the intergenerational feature of nonideal justice requires clarification.

In order to determine what should be done under a particular set of nonideal conditions at a particular time $t$, it is not sufficient to consider the interests of people who live at $t$. This is because, what is done at $t$ does not just have immediate effects: it also has long-term causal implications, impacting both what can and needs to be done at future times. For instance, if at $t$ policies are adopted that have short-term benefits but, long-term, impose severe costs, then different people in the future will have to endure the burdens of such costs. Consequently, the contractualist procedure must model considerations of intergenerational, as opposed to merely intragenerational, justice.

---


11 Ibid. 118. Mills notes that in *A Theory of Justice*, race is not initially listed as one of the features that you do not know about yourself behind the veil of ignorance. However, in *Political Liberalism* this omission is rectified. (See: Mills, “Rawls on Race/Race in Rawls,” 164-165.) I think it is fair to say that even in *A Theory of Justice*, the fact that features such as race are hidden by the veil is obviously implicit.


13 Ibid. 118.
Of course, to forestall a potential objection, what is done in a particular set of nonideal conditions does not necessarily have causal implications for what different future people under different nonideal conditions can and need to do. Under some empirical conditions—very fortunate empirical conditions—it will be possible to achieve ideal justice in one generation. My point is just that ideal justice could—and I strongly suspect in practice almost always will—take at least several generations for any actual state (e.g., the present day US) to achieve. Therefore, it has to be shown that nonideal principles of justice can coherently and fairly be applied in an intergenerational context. Even if an actual society is in the extraordinarily fortunate position of being able to achieve ideal justice in a single generation, the same nonideal principles of justice that I identify and defend below apply.

Rawls briefly discusses intergenerational justice in *A Theory of Justice*. He argues that the contracting parties should be modeled as knowing that they all belong to the same generation, although as having a *concern* for their immediate descendants. In the literature this is often described as the “present time of entry” model of intergenerational justice. During his long career, Rawls’s conception of the present time of entry modeled evolved. Furthermore, Rawls’s various ways of modeling intergenerational justice have been extensively criticized.

So as not to break up the flow of this chapter—and digress too far into an extensive literature review of a comparatively obscure topic—I will confine myself to a statement of how I think the intergenerational feature of nonideal justice should be modeled. I defend

---


this model, and break with other prominent ways of modeling this intergenerational feature in Appendix I. A single party ignorant of which generation or social position within a particular generation they will occupy is required to select principles of justice that rationally maximize their interests under conditions of uncertainty. Given that the contracting party is ignorant of which generation they will actually be born in, it is rational for the party to select principles that are intergenerationally impartial—hence fair—in the requisite sense. I call this the “representation of all generations model.”

The claim that there is only a single party requires brief clarification. In a Rawlsian contractualist framework it has become conventional to refer to multiple parties. However, given that all the parties would be situated in exactly the same way, a single party would be sufficient.17 (Although strictly speaking a single party would be sufficient, in what follows I will retain the stylistic convention of referring to multiple parties.)

2.5. Preliminary assumptions and distinctions

I will derive a set of nonideal principles of justice by posing a question to the nonideal contracting parties. I will begin by clarifying some assumptions and distinctions upon which this question depends.

2.5.1 An assumption of strict compliance

The nonideal contractual parties—exactly like the ideal contracting parties—select principles of justice against the backdrop of the assumption that all actual people will strictly comply with the nonideal principles of justice that they select. It is important to clarify this assumption because, particularly in the context of nonideal theory, it potentially invites misunderstanding. One of the causal reasons that actual societies are in nonideal conditions is that people fail, or historically have failed, to strictly comply with the demands of justice. This fact does not, however, entail that the contracting parties cannot assume that actual

people will strictly comply with the nonideal principles of justice that they select to respond to such injustice.

This assumption is not realistic but that is not a problem. The central justification for this assumption is the value of the conception of justice that it generates: it allows the contracting parties to select principles of nonideal justice that specify what justice *simpliciter* requires, without that selection being tainted by actual people’s expected noncompliance. I will defend the value of such a conception of justice in chapter 3, by arguing that it is necessary for justice to perform what I term the “agent-guiding function of justice”.

A significant consequence of my approach is that certain types of questions that might plausibly be thought to be the focal point of nonideal theory are temporarily left unaddressed. In particular, some further account is required for how principles that presuppose strict compliance are valuable in nonideal conditions in which people do not, as

---

18 In the context of ideal theory, A. John Simmons presents a different justification for the assumption of strict compliance. He writes that ‘[i]f instead we try to evaluate principles in terms of how societies governed by them would operate with a “normal” amount (or a certain percentage) of noncompliance with them (supposing we can even make sense of that hypothesis), we will likely find both that our evaluations yield quite indeterminate results and that the results depend on more than simply the different ordering effects of the principles being compared.’ (A. John Simmons, “Ideal and Nonideal Theory,” *Philosophy & Public Affairs* 38 (2010): 5-26, pp. 8-9.) (Cf. Laura Valentini, “On the Apparent Paradox of Ideal Theory,” *Journal of Political Philosophy* 17 (2009): 332-355, p. 339.) I am neutral about whether this alternative justification succeeds. I do, however, have a couple of reservations. First, it is not clear to me why relaxing the assumption of strict compliance would yield results that are indeterminate. Just as a professor could devise a policy for attendance against the backdrop of an assumption that her students will partially comply with such a policy, so it seems to me the contracting parties could try to determine principles that would maximize their prospects, against the backdrop of an assumption that actual people will only partially comply with the principles selected. (To be sure, any specific level of non-compliance would be arbitrary: there would, for instance, be no reason to adopt a rate of 6% non-compliance over that of a rate of 7% non-compliance. However, this does not mean that the results would be indeterminate. Furthermore, it could be argued that even if the selection of a rate of non-compliance is somewhat arbitrary, it is still preferable to an assumption of strict compliance because it is more in line with how actual people behave.) Second, relaxing the assumption of strict compliance would make the principles selected relative to a rate of non-compliance. However, it is not clear—as Simmons implies—that this is problematic. As I noted above the parties have a comprehensive knowledge of social science so the principles will be relative to a number of facts. As long as the rate of non-compliance—like these other facts—is held fixed when comparing different candidate principles, then comparative judgments about which candidate principles are preferable seem as though they will be intelligible.
a matter of fact, strictly comply. I will provide this account in chapter 3. However, in this chapter I will undertake the initial task of deriving a set of abstract principles of nonideal justice that assume strict compliance.

2.5.2 An assumption that Rawls’s ideal theory is correct and that a society is relatively affluent

In selecting principles, the parties also assume that the ideal theory of justice that Rawls defends in *A Theory of Justice* is correct. As I will explain below, some of my analysis is relevant even if this presupposition is false. Furthermore, as noted in chapter 1 (§1.5.1), my nonideal theory of justice presupposes the economic conditions of relative affluence. In the final section of this chapter I will, however, briefly speculate about the content of ideal theory if this economic condition does not obtain.

2.5.3 Two preliminary distinctions

It is important to make two related distinctions that are salient in nonideal conditions but not in ideal conditions. Both distinctions can be illustrated using the same real world example. Under Mayor Giuliani's tenure as mayor of New York, the New York City Police Department increasingly implemented a so-called “Stop-and-frisk” practice. As the name of this practice suggests, it allowed police officers to stop, question and frisk pedestrians under a standard of reasonable suspicion.\(^19\) This practice—particularly given that the standard of “reasonable suspicion” was so vague that it could be interpreted however the officers wanted—clearly violated the basic liberty of freedom of the stopped person. This is because, as Rawls notes, this basic liberty includes freedom from psychological oppression.\(^20\) Quite understandably given its nature, the practice induced a great deal of fear, which prevented innocent citizens from exercising their freedom in a public space.

The first distinction to note is between the state’s official recognition of basic liberties (e.g., in documents such as a written constitution) and the degree to which people

\(^19\) N.Y. Criminal Procedure Law 140.50.

are in fact able to exercise their basic liberties because of practices such as stop-and-frisk, and systematic non-compliance with the demands of justice by private citizens.

The caveat of “systematic” non-compliance is important. If actual people’s ability to exercise their basic liberties was just threatened by very isolated instances of actual people’s noncompliance with the demands of justice, then such instances would plausibly fall outside the scope of a political conception of justice. To put the point more concretely, suppose something like racism was merely an idiosyncratic prejudice—comparable to a prejudice that a couple of individuals have against people with the proper name “Kyle.” If this were the case, then the phenomena of racism would appear to lie outside the scope of political justice but (plausibly) within the scope of ethics. It is worth pausing to note that this analysis shows that actual people’s non-systematic non-compliance with the demands of justice is compatible with the realization of ideal justice, for such non-compliance lies outside the scope of justice.

A consequence of this first distinction is that a state can be in nonideal conditions even if its official recognition of basic liberties conforms to the ideal Liberty principle—for people’s ability to exercise their basic liberties may, nonetheless, not conform to the ideal Liberty principle.

The second, related distinction is that within nonideal conditions different people may be able to exercise their basic liberties to different degrees. For example, 90% of the people who were stopped and frisked in New York City were black or Latino and had

---


22 As Thomas Pogge notes, there are a couple of instances in Rawls’s work in which he (at least implicitly) recognizes this first distinction. For instance in “Social United and Primary Goods,” Rawls writes: ‘Whether the basic structure guarantees equal liberty of conscience, or freedom of thought, [i.e., liberties that are classified as basic by Rawls] is settled by the content of the rights and liberties defined by the institutions of the basic structure and how they are actually interpreted and enforced.’ (John Rawls, “Social Unity and Primary Goods, in *Utilitarianism and Beyond*, ed. Amartya Sen and Bernard Williams (Cambridge: Cambridge University Press, 1982), 159-186, p. 163 (emphasis added). (Also cited by Thomas Pogge, “Three Problems with Contractarian-Consequentialist Ways of Assessing Social Institutions,” *Social Philosophy and Policy* 12 (1995): 241-266, p. 256.)
committed no crime.\textsuperscript{23} Due to the stop-and-frisk practice, black and Latino New Yorkers were \textit{ceteris paribus} less able to exercise their basic liberties than other citizens, because of the direct effects of this practice or the fear that it induced. In some nonideal conditions, then, we can distinguish between the different degrees to which people are able to exercise their basic liberties.\textsuperscript{24}

In order to develop an adequate theory of nonideal justice it is necessary to draw these distinctions. Without them, such a theory can offer no account of what ought to be done when there is a disparity between the liberties that the state officially recognizes and citizens’ actual ability to exercise such liberties. If such a theory could not offer such an account, it would be blind to an important type of injustice. It would, thereby, be vulnerable to the familiar, broadly Marxist, objection to traditional liberal theory: liberalism consists in a set of rights and liberties that are mere abstractions, which are completely insensitive to the deeper underlying social reality.\textsuperscript{25}

In order to ward off misunderstanding one point is worth clarifying. A distinctive feature of Rawls’s ideal theory is that, with the exception of the political liberties, he denies that equal worth of liberty (i.e., usefulness to persons of their liberties) is a requirement of justice.\textsuperscript{26} Rawls thinks that such a requirement would be unfair and socially divisive.\textsuperscript{27} After all, as he points out, some people may think it their religious duty to build magnificent cathedrals:

\begin{footnotesize}
\begin{enumerate}
\item[24] In \textit{A Theory of Justice} Rawls also makes such a distinction. This is because he recognizes that in ideal conditions the content of basic liberties can be determined with respect to a single social position of equal citizenship, while in nonideal conditions ‘…other positions may need to be taken into account. If, for example, there are unequal basic rights founded on fixed natural characteristics, these inequalities will single out relevant positions.’ (Rawls, \textit{A Theory of Justice}, 84.)
\item[27] Following Samuel Freeman, \textit{Rawls} (Oxford: Routledge, 2007), 62.
\end{enumerate}
\end{footnotesize}
To guarantee the equal worth of religious liberty would mean, then, that society is to devote social resources to these citizens rather than to others whose understanding of their religious duties calls for far fewer material requirements.28

It is important to note that neither of the distinctions that I am drawing in the context of nonideal theory need be unattractive to philosophers who accept Rawls’s ideal theory. For the same reason he supplies, I think that Rawls is correct to conclude that justice need not require that people actually enjoy equal worth of liberty. However, if due to state based injustice or actual people’s noncompliance with the demands of justice, some citizens not even capable of exercising their basic liberties to an equal extent, then this is a failing of justice that a nonideal theory of justice must be appropriately responsive to.

An analogous set of distinctions also apply with respect to fair equality of opportunity. Regardless of what the law officially recognizes, some people may not actually have access to fair equality of opportunity in an employment or educational context. Furthermore, because of, for instance, racial discrimination against a particular group, different people may have access to fair equality of opportunity to different degrees.

2.6 Presenting and defending a set of nonideal principles of justice

Having set out the necessary preliminaries I can proceed to derive a set of nonideal principles of justice.

2.6.1 Question posed to the nonideal original-position contractors

The nonideal contracting parties are asked to select nonideal principles of justice, which rationally maximize their own interests in nonideal conditions, given that they do not know which social position they will occupy or in which generation they will be born. They assume that actual people will strictly comply with the nonideal principles of justice that they select, that Rawls’s ideal theory of justice is true, and that the society into which they will be

born is relatively affluent. Crucially, they also know that regardless of what set of basic liberties the state officially recognizes, actual people’s ability to exercise their basic liberties will not satisfy Rawls’s ideal Liberty principle. Furthermore, it is possible that different people will be able to exercise their basic liberties to different degrees. Analogous points also apply with respect to fair equality of opportunity.

2.6.2 The nonideal principles of justice selected

I will begin by presenting the (admittedly rather baroque) set of nonideal principles of justice that I think the contracting parties would settle on. I will then outline the reasoning that would lead to this selection.

The First Principle: “Nonideal Liberty”

In this principle, “comparative advantage” or “disadvantage” is with respect to an actual ability to exercise one’s basic liberties.

I: Measures should be instituted to realize Rawls’s ideal Liberty principle of justice: ‘Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all.’ 29 The measures that are both permissible and required in order to achieve this aim are specified by the following clauses.

II: In determining what measures should be implemented in a particular set of nonideal conditions at time t, it is not sufficient to consider what measures would be the most effective means of increasing the standing of comparatively disadvantaged people at t. Rather, measures should be introduced that are the most effective means of increasing comparatively disadvantaged people’s standing prior to the full realization of ideal justice, regardless of which generation they are born in.

29 Rawls, A Theory of Justice, 266.
III: If comparatively disadvantaged people are disadvantaged to different degrees, priority should be given to the least advantaged.

IV: Comparatively disadvantaged people’s standing should be increased in such a way that it imposes as few demands on comparatively advantaged people as possible.

V: If it is not possible to increase the standing of disadvantaged people without imposing costs on comparatively advantaged people, such costs should be distributed according to the following two principles: (1) if possible, costs should be imposed on comparatively advantaged people;30 (2) the relative significance of these costs must be determined by the priority ordering of the ideal principles of justice (e.g., the basic liberties have priority over equality of opportunity).

The first part of the second principle: “Nonideal Fair Equality of Opportunity (FEO)”

Measures should be instituted to realize Rawls’s ideal FEO principle. The measures that are both permissible and required in order to achieve this aim are specified by the above clauses II-V. In such clauses, “comparative advantage” or “disadvantage” is with respect to an actual ability to access fair equality of opportunity.

The second part of the second principle: “a nonideal intergenerational different principle”

Measures should be instituted to realize Rawls’s ideal difference principle. The transition towards this difference principle should, itself, be arranged in terms of a nonideal intergenerational difference principle: economic inequalities in this

30 Assuming, of course, that such costs are not so great that bearing these costs would make people who were antecedently comparatively advantaged more disadvantaged than people who were antecedently comparatively disadvantaged. If people are equally advantaged qua their ability to exercise their basic liberties, costs should be imposed on people who are comparatively advantaged qua their ability to have fair equality of opportunity. If they are also advantaged equally in this respect, then costs should be imposed on people who are comparatively advantaged qua their index of primary economic goods.
transition are acceptable if and only if they benefit the least advantaged, regardless of which generation prior to the full realization of ideal justice the least advantaged people are born.

*Priority Rule 1:*

These principles are to be arranged in serial order with the first principle prior to the second. The first part of the second principle has priority over the second part.

*Priority Rule 2:*

A fair realization of ideal justice (as specified by the two ordered nonideal principles of justice) has priority over the most efficient realization of ideal justice (i.e., transitional paths that violate the requirements of nonideal justice—however efficient and effective—are impermissible).

2.6.3 The basic reasoning leading to the selection of the nonideal principles of justice

It is immediately striking that this set of nonideal principles of justice is intimately connected with Rawls’s ideal principles of justice in two different ways. First, the ultimate aim of all the nonideal principles of justice is to realize different components of ideal justice. Second, the transitional path that they specify to such ideals has similar content to Rawls’s ideal principles of justice. This result is significant. After all, as I cautioned in Chapter 1 (§1.4.2), establishing something about the content of ideal justice does not straightforwardly entail anything about the content of nonideal justice.

I will begin by defending each clause of the nonideal Liberty principle. The contracting parties would select clause I because they want to advance their ends as effectively as possible. In particular, they want to realize the ideal Liberty principle so that they can accrue the benefits of living in a more just world, in which they have an ability to exercise their basic liberties to a greater extent. The same basic set of considerations and values inform the construction of nonideal principles of justice as ideal principles of justice. This is not merely because the ideal and nonideal original positions are set up in such a
similar way, but also because the same considerations that dictate how a perfectly just society should be arranged inform how the parties want to transition to that perfectly just society. Just as the ideal contracting parties want to maximize their ability to exercise their basic liberties in ideal conditions regardless of which position they end up occupying, so the nonideal contracting parties want to maximize their ability to exercise their basic liberties in nonideal conditions regardless of which position they end up occupying. Furthermore, because they know that they could end up in a position in which their ability to exercise their basic liberties will be restricted they would endorse measures to ameliorate such restrictions.

Clause II is selected because the parties do not know when they will be born. Therefore, they do not want measures to be instituted at a particular time that just improve the standing of the comparatively disadvantaged at that time; consequently, they reject measures that privilege the interests of a particular generation prior to the full realization of ideal justice because they do not know in which generation they will be born.

Clause III is chosen because given the choice problem posed, it is rational for the nonideal contracting parties—like the ideal contracting parties—to instigate measures to insure that their social standing is as good as possible if they end up occupying the position of the least advantaged.\footnote{Like most aspects of Rawls’s philosophy, there is a vast literature discussing both why and whether it is rational for the parties to favor the interests of the comparatively disadvantaged. I will not, here, attempt to weigh in on the debate or to further defend Rawls’s position. For a good overview of both the basic contours of this debate, and how Rawls’s conception of rational choice within his contractualist framework evolved throughout his writing career, see: Gerald Gaus and John Thrasher, “Rational Choice in the Original Position: The (Many) Models of Rawls and Harsanyi,” in \textit{The Original Position}, ed. Timothy Hinton (Cambridge: Cambridge University Press, 2016), 256-291.}

Clause IV is selected because although the parties are prepared to impose costs on comparatively advantaged people for the sake of improving the social standing of the least advantaged, they are not indifferent to the nature of these costs. After all, they could end up occupying the social position of comparatively advantaged people. Therefore, \textit{ceteris paribus} they would prefer for the costs that fall on comparatively advantaged people to be as small as possible.

Clause V is selected because the parties recognize that, if some costs or burdens are necessary to improve the standing of the comparatively disadvantaged, it is preferable for...
such costs to be borne by the comparatively advantaged. The reasoning for this is straightforward: given that the parties do not know which social position they will occupy, they prefer to impose costs on the comparatively advantaged so long as this increases the standing of the least advantaged and does not bring the overall new standing of the comparatively advantaged down to a level below the new standing of the comparatively disadvantaged. As I noted above, their reasoning is not guided by utilitarian principles; rather it is rational for them to aim to improve the lot of the least advantaged. The significance of the costs is determined by the priority ordering of the ideal principles of justice because, as I also noted above, the same basic set of considerations and values inform the construction of nonideal theory as ideal theory.

Precisely the same set of reasons also apply \textit{mutatis mutandis} to the selection of the nonideal FEO principle, the content and structure of which is symmetrical to that of the nonideal Liberty principle. Interestingly, the nonideal intergenerational difference principle is more concise than the other principles. This is because the realization of the intragenerational ideal difference principle can itself be elegantly captured by a nonideal intergenerational difference principle. Finally, the priority rules are selected because the same reasons that motivate the ordering of ideal principles of justice also motivate the realization of such principles: basic liberties are, for instance, valued over fair equality of opportunity, and justice is valued over efficiency.

2.7 The relationship between ideal and nonideal theory

The selection of this set of nonideal principles of justice constitutes the relationship between ideal and nonideal theory. This relationship is transitional: the goal of ideal justice is realized \textit{via} the transitional path specified by the nonideal principles of justice.

2.7.1 Deepening the justification for this transitional conception

The contractualist framework can be employed in a different way to deepen the justification for this transitional conception. We can imagine the contractualist parties
entertaining other candidate conceptions of the relationship and consider why they would reject them.  

A teleological conception: nonideal theory should serve as a political technology for the effective realization of ideal justice

On this conception of nonideal theory the only type of question that arises in nonideal conditions is one of efficacy: how can ideal justice be most directly realized in the future.

I think that it is natural to slip into this way of thinking: an ideal conception of justice is after all a conception of how we think a perfectly just society should be structured; consequently, it can present itself as having a pressing practical claim on us to such an extent that the only question that seems pertinent in nonideal conditions is the technical question of how we can effectively bring about this ideal conception of justice. Indeed, as the history of Shining Path communist party in Peru dramatically illustrates, this conception sometimes has a deep grip on actual political practice.

However, at least within the nonideal contractualist framework, this conception of nonideal theory would be strongly rejected. The parties do not know when they will be born. Consequently, they will not assent to this conception because it could impose huge burdens on them for the sake of different future people. Consequently, they would endorse principles that constitute *sui generis* considerations of nonideal justice in nonideal conditions.

---

32 This use of a contractualist framework is comparable to Rawls’s own theorizing in *The Law of Peoples*. Rather than being tasked with determining a set of principles *ex nihilo*, the contracting parties are required to consider—and either assent to or reject—a set of traditional principles drawn from international law. (See: John Rawls, *The Law of Peoples* (Cambridge Mass.: Harvard University Press, 2001), 36-37.) Analogously, in my analysis above, the contracting parties are asked to consider and either assent to or reject candidate conceptions of the relationship between nonideal theory and ideal theory.

An approximation conception: under nonideal conditions, nonideal theory directs us to approximate ideal justice

As I explained in Chapter 1 (§1.4.4), Rawls himself sometimes conceives the relationship between ideal and nonideal theory in precisely this way: under each and every set of nonideal conditions we should try to make our actual institutions mirror—presumably as fully and effectively as we can—what a realization of ideal principles of justice would require.34

The nonideal contracting parties would reject this conception of the relationship between ideal and nonideal theory. Given that they want to maximize their own prospects they would not attach any intrinsic value to approximating ideal principles of justice under any particular set of nonideal conditions. In particular, they would want to avoid a more direct approximation of ideal justice in a particular set of nonideal conditions if this sets back the interests of different comparatively disadvantaged people in different nonideal conditions. This is because they could end up occupying the position of such a comparatively disadvantaged people. Rather, they would only want to approximate ideal justice in a particular set of nonideal conditions to the contingent extent that this was part of the best long-term program for realizing ideal justice, as specified by the above nonideal principles of justice.35

---


35 There are additional, more technical, reasons for rejecting this conception of the relationship between ideal and nonideal theory. As Robert Goodin explains: ‘The second-best state of affairs is not necessarily one in which your ideal conditions are realized more rather than less completely.’ (Robert E. Goodin, “Political Ideals and Political Practice,” British Journal of Political Science 25 (1995): 52). This was an argument that was famously made by Lipsey and Lancaster against Pareto optimality. Goodin gives the following, more intuitively accessible illustration of their central point: ‘Your ideal car, let us suppose, would be a new silver Rolls. But suppose the dealer tells you none is available. The point of the general theory of the second best is this: it simply does not necessarily follow that a car that satisfied two out of your ideal cars three crucial characteristics is necessarily second best. You may well prefer a one-year-old black Mercedes (a car unlike your ideal car in every respect) over a new silver Ford (which resembles your ideal car in two out of three respects).’ (Ibid. 53) Therefore, simply trying to approximate the ideal under nonideal conditions—in which the ideal cannot be fully realized—will not in general be a reliable guide for determining what we should do under nonideal conditions. (Cf. R. G.
An exemplification relation: in nonideal conditions we should act as we would if ideal conditions fully obtained

In the following passage from *The Groundwork of the Metaphysics of Morals* Kant suggests that this is the appropriate conception of how ideal theory applies to nonideal conditions:

[E]very rational being must so act as if he were through his maxim always a legislating member in the universal kingdom of ends.\(^{36}\)

The kingdom of ends is just a regulative ideal but the proper moral relation to this ideal—even under nonideal conditions—is one of making yourself a fitting member (as both a legislator and a citizen, to use familiar Kantian terminology). If this Kantian claim is correct then nonideal theorizing is simple: under nonideal conditions do exactly as you would do under ideal conditions.\(^{37}\)

---


\(^{37}\) As Christine Korsgaard notes this Kantian conception of the demand to adhere to ideal principles under nonideal conditions—regardless of the consequences—is tied to Kant’s religious views. As she explains: ‘Kant is by no means dismissive towards the distressing problems caused by the evil conduct of other human beings and the unfriendliness of nature to human ideals, but his solution to these problems is different. He finds in them grounds for a morally motivated religious faith in God. Our rational motive for belief in a moral author of the world derives from our rational need for grounds for hope that these problems will be resolved.’ (Christine M. Korsgaard, *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press, 1996), 149.) Korsgaard also notes that elsewhere in Kant’s corpus, such as in his *Lectures on Ethics*, we sometimes encounter a less pure and strict conception of the way in which we ought to adhere to moral ideals (i.e., people’s noncompliance with moral demands is construed as a sufficient reason for altering the demands that compliant agents face in conditions of partial compliance). (See: Ibid. 154). For instance Kant states: ‘But as men are malicious, it cannot be denied that to be punctiliously truthful is often dangerous…if I cannot save myself by maintaining silence, then my lie is a weapon of defense.’ (Immanuel Kant, *Lectures on Ethics*, trans. Louis Infield (U.S.A. Hackett, 1980), 228.)
In response, it is first useful to draw a distinction between the obligations that fall on an individual in nonideal conditions, and the question of how the basic structure of society should be altered under nonideal conditions. It seems as least possible for me qua individual to act, even in nonideal conditions, in such a way that would be required of me if I were a legislator and citizen in a kingdom of ends. I can fulfill, in other words, the individual moral responsibilities that I would face in an ideal world—in which everyone was doing the right thing—in nonideal conditions. Of course, because other people are not all doing the right thing I may, as Kant’s notorious response to Benjamin Constance in “A Supposed Right to Lie for Philanthropic Concern” illustrates, have to endure unusual burdens or face painful consequences.\textsuperscript{38}

In contrast, it seems difficult to understand how political actors—who are employed by institutions that constitute the basic structure of society—could exemplify ideal justice under nonideal conditions. This is because the just action of a political agent under nonideal conditions is not sufficient to partially exemplify ideal justice. Rather, the basic structure of political institutions must actually be just (i.e., governed by just practices or “rules of the game”). For instance, a Supreme Court Justice cannot act as she would if ideal conditions obtained under nonideal conditions because her actions would not have the requisite institutional backing. That would only be possible if ideal justice was realized.

However, even if these initial concerns about it even being possible for political actors to exemplify ideals can be assuaged, it is clear that the contracting parties would reject a conception of nonideal theory as one in which we should try to exemplify ideals. The contracting parties, after all, care about maximizing their individual prospects. They want to accrue the benefits of living in as just a society as possible, whilst avoiding bearing disproportional costs on the road to realizing ideal justice. They attach no value to the piecemeal exemplification of components of ideals of justice under nonideal conditions. The conception of nonideal theory as one in which we should try to act as we would if an ideal world obtained is one that is (at most) appropriate for individual moral responsibilities under nonideal conditions.\textsuperscript{39}


\textsuperscript{39} Of course it might not even be an appropriate conception of what individuals should do under nonideal conditions. As I explained following Korsgaard, in footnote 37 Kant himself—the proclaimed “poster
As I noted in chapter 1 (§1.4.4) Korsgaard argues that Rawls’s nonideal theory is directly constrained by features of his ideal theory. Tamar Schapiro and Robert S. Taylor have done important work clarifying and elaborating the details of Korsgaard’s interpretation. Schapiro argues that a deontological nonideal theory, such as that of Rawls or Kant, must be consistent with the spirit of ideal theory even if it can violate the letter.

I think that the following is an interesting test case for further exploring the general topic of how moral obligations under conditions of partial compliance relate to what would be required of agents in conditions of full compliance—a type of case that has not been explored by contemporary Kantians or Murphy: a professor, of an illustrious academic institution such as Yale, takes a firm stance against grade inflation: “Ideally grades should be informative and the average grade should be a C+ rather than an A- so I will not give in to grade inflation but instead grade my students as they should be graded.” Note that there is a concern that the Professor’s commitment to what we might refer to as “ideal integrity” is misplaced. In nonideal academic conditions there is in fact rampant grade inflation and by simply ignoring this fact and doing what should ideally be done under nonideal conditions the professor is unfairly treating her students who will be harshly judged in comparison to the students of other professors who have “given in” to grade inflation. As a final remark, I would tentatively suggest that if we want to explore how the demands that individuals ought to face in conditions of partial compliance relate to the demands that they would face in conditions of partial compliance, it will be helpful to distinguish between different types of moral demands. Simply put, we could coherently think that some moral demands are the same in conditions of partial and full compliance and others are different.


Taylor adds that it must be constrained by Rawls’s general conception of justice, and it must reflect the priority relations of ideal theory in its “order of action.”

There is, however, no reason to think that the contracting parties would insist on nonideal theory being constrained by features of ideal theory unless it maximized their own interests.

Perhaps in response, Korsgaard and her followers could object that this was a reason to reject my approach of determining the content of nonideal theory in a contractualist framework: nonideal theory must be constrained by features of ideal theory, in order to preserve its deontological character.

I think that this concern can be assuaged by noting that the deep motivation for Korsgaard’s approach is also captured in my alternative approach. Taylor, for instance, argues that unless Korsgaard’s interpretation is adopted nonideal theory would:

…consist of nothing but a goal (ideal conditions) and a pair of conditions following immediately from it (provisionality and instrumentality), [therefore] it would appear to be consequentialist…to our question of how we should achieve ideal conditions, the nonideal theory would seemingly answer, “by any means necessary.”

However, this argument depends on the construction of a false dichotomy. It is not the case that either nonideal theory is constrained by features of ideal theory and consequently it is deontological, or it is not constrained by features of ideal theory and consequently it is crudely consequentialist. There is an alternative possible conception of the relationship that I have provided a detailed contractualist defense of: the transition to ideal justice is subject to constraints of fairness and not merely concerned with realizing the goal of ideal justice as directly as possible. However, it is subject to such constraints of fairness because it is guided by *sui generis* nonideal principles of justice, rather than being constrained by features of ideal theory.

---

43 Ibid. 147-51.
2.7.2 Why Rawlsian nonideal theory is exclusively forward-looking

A significant feature of the relationship between nonideal theory and ideal theory is that the nonideal principles of justice are exclusively forward-looking: they specify how we should fairly transition to ideal justice. The reason they are selected is to bring about good consequences in the future; they impose no demands of rectification because of injustice in the past. Why are the principles of nonideal justice purely forward-looking? The reason, essentially, is that given the choice problem posed it is rational for the parties to select forward-looking principles, for such principles allow them to advance their own ends as effectively as possibly. It is impossible to change the past and, therefore, in selecting a set of principles that apply to a particular set of nonideal conditions at $t$, they are exclusively concerned with how the principles improve the situation of people in the present and future. They do not care about the rectification of past injustice for its own sake and, therefore, reject backward-looking principles that stipulate that the rectification of past injustice requires certain things, regardless of whether it improves the individual prospects of people in the present and future.

---

44 I use the term “forward-looking” to refer to the claim that the reason rectification for historic injustice should be sought is to bring about good consequences in the future; “backward-looking” to refer to the claim that the reason rectification for historical injustice should be imposed in the present and future is to redress injustice in the past.

45 Following Simmons, “Ideal and Nonideal Theory,” 33.

46 Pogge argues that although Rawlsian contractualism and consequentialism are positions that are standardly contrasted with one another in contemporary political philosophy, they are actually far more similar than is typically understood. As he explains: ‘Understood as guides to the assessment of social institutions, contractarianism and consequentialism are for the most part not competitors but alternative presentations of a single idea: both tend to assess alternative institutional schemes exclusively by how each would affect its individual human participants.’ (Pogge, “Three Problems with Contractarian-Consequentialist Ways of Assessing Social Institutions,” 246.) (Cf. Derek Parfit, On What Matters (Oxford: Oxford University Press, 2011), 354). I think that it is difficult to assimilate Rawls’s position to consequentialism, for Rawls famously thinks that the term “consequentialism” was ambiguous and insisted instead on distinguishing between theories that were teleological as opposed to deontological. (See Samuel Freeman, “Consequentialism, Publicity, Stability, and Property-Owning Democracy,” in Justice and the Social Contract: Essays on Rawlsian Political Philosophy (New York: Oxford University Press, 2006), 75-110, for an illuminating discussion of this important point.) Nevertheless, I agree with both Pogge and Parfit that the contractors “care about getting the right results”,


In the limited contemporary Rawlsian literature on the topic, the claim that Rawls’s theory of justice is purely forward-looking has occasionally been challenged. Moises Vaca, for instance, develops a reading of Rawls in which she argues that the full institution of Rawlsian ideal justice requires—at least at the level of legislation—the full rectification of historic injustice. She writes:

According to Rawls, one of the liberties that must be equally distributed by the institutions of the basic structure is the ‘freedom of the person which includes freedom from psychological oppression’. Because of the tremendous effects of non-rectified past political violence on victims’ psychological life, it is hard to see how such a liberty would be distributed equally amongst victims and non-victims without institutional provisions to come to terms with the past.

Vaca adopts a similar strategy with respect to the value of self-respect. She correctly notes that Rawls construes this as perhaps the most important primary good and she argues that given core facts about human psychology a rectification of historical injustice is necessary for achieving self-respect in the present.

The first crucial thing to note is that Vaca’s defense of the rectification of historical injustice is mediate. Her claim is not that justice simpliciter requires the rectification of even though—in contrast to familiar Utilitarian positions—they insist that the right results consist, at least in part, in deontological restraints.


Ibid. 308, quoting Rawls, A Theory of Justice, 53. (She also refers to the following passages in Rawls’s corpus: Justice as Fairness a Restatement, 44; Political Liberalism, 291.)

See Rawls, A Theory of Justice, 386.

historical injustice. Rather, as a contingent matter of fact the implementation of Rawlsian principles of justice—which contain no essential backward-looking components—requires us to respond appropriately to historic injustice.

However, once this is appreciated, it is easy to show that the reason for compensation in the present and future is actually, on careful inspection, purely forward-looking: compensation should be provided to victims of historic injustice in order to remove the unpleasant present and future psychological consequences of such injustice. The reason for such institutions is not the backward-looking one that this is what is required in order to compensate for historic injustice.

A comparison between backward and forward-looking conceptions of victim compensation in the context of punishment further illuminates this point. A purely forward-looking consequentialist approach can provide a justification for the institution of victim compensation. However, the reason it provides for having this institution is—like Vaca—the purely forward-looking one that it is necessary to bring about positive effects in the future, rather than the retributivist backward-looking claim that compensation is required because of past wrongdoing.51

In recent work, Simmons argues that the Rawlsian approach involves a kind of blindness to history, and does not even leave conceptual room to make sense of an important class of morally compelling considerations viz. historical injustice.52 Simmons’ charge, potentially invites misunderstanding. As Vaca’s work illustrates, Rawlsians are not blind to historical injustice in the sense that they are indifferent to it; there are reasons to rectify past injustice, although these reasons are purely forward-looking. However, Simmons is clearly right that the Rawlsian contractualist approach is not able to make sense of

---

51 See David Boonin, The problem of Punishment (Cambridge: Cambridge University Press, 39-41), for a further clarification of this point.
52 See: A. John Simmons, Boundaries of Authority (Oxford: Oxford University Press, 2016), Chapters 2 and 3. Simmons does not just object to the Rawlsian approach. His arguments are presented against what he terms “Kantian functionalism” in general. Kantian functionalists, who include Rawls and contemporary Rawlsians, advance substantively different normative claims but they are united by their “structuralist” approach to justice: to the claim that falling under just institutions at a particular time is sufficient to confer full legitimacy on the state.
historical wrongness *simpliciter* (e.g., as historical wrongness because it involves the historic violation of peoples’ rights).

The Rawlsian approach is, consequently, unable to accommodate the intuition that historical injustice is something in and of itself that is wrong and should be rectified regardless of the effects that it has on victims’ psychological lives in the present. Given this fact it could be objected that Rawls’s way—and by extension my way—of excluding a backward-looking component is question-begging: the original position is set up in such a way that it is rational for the parties to select forward-looking principles of justice. However, no argument is supplied for modeling the original position in such a way.

This approach might seem particularly problematic given a commitment to a reflective-equilibrium methodology: many people, for instance, who advocate on behalf of Black and Native Americans argue that the reason that compensation is owed from the United States government is not because this will ameliorate and facilitate the process of future political development; rather they point to the fact these minority groups were the victims of historical injustice and this is what justice *simpliciter* requires. Therefore, perhaps, the principles of nonideal justice that I have defended cannot stand in a reflective-equilibrium with our considered intuitions.

In response, I think that it should be acknowledged that my Rawlsian approach is somewhat question-begging at least to the extent that there is no direct argument against a backward-looking approach. However, the crucial question is whether a backward-looking or forward-looking account is ultimately more theoretically powerful and plausible. In chapter 5 I will highlight the ability of my forward-looking account to offer a justification of non-meritocratic affirmative action. The question of whether a forward or backward looking approach is ultimately more plausible lies outside of the scope of my project: it would require developing both approaches and assessing all of their comparative costs and benefits. However, there is no reason in advance to suppose that my Rawlsian approach is untenable. Indeed, if it turns out to be the more powerful and plausible approach it could supply a reason for people to modify some of their antecedently backward-looking intuitions.

---

2.8 A refinement and clarification of Rawls’s conception of nonideal theory

I am now in a position to explain how my view overcomes the problems and deficiencies with the sketch of nonideal theory that we inherit from Rawls.

2.8.1 The content and structure of nonideal theory

Most obviously, the nonideal principles of justice fill out the normative content of the indeterminate permissibility condition that we inherit from Rawls. Furthermore, given that they are structured by priority rules there is no methodological tension between ideal and nonideal theory.

Recall that in Chapter 1 (§1.4.2) I argued that a central normative question that Rawls is silent on is whether we can violate considerations that animate the selection of ideal principles of justice in order to effectively transition towards ideal justice in the long-term. The nonideal principles of justice that I have derived allow me to address this question. The answer is complex. On the one hand the nonideal principles of justice are animated by the same considerations that animate the selection of ideal principles of justice. Consequently, they prohibit using people as a mere means to realize ideal justice in the future. However, they permit certain costs to fall on people—that would not be permissible in ideal conditions—if these costs constitute a fair distribution of the burdens of transitioning to a more just world. As I will explain in chapter 5, in some nonideal conditions these burdens can be quite extensive as they can be used to justify non-meritocratic affirmative action policies.

2.8.2 The relationship between ideal and nonideal theory

As I noted in chapter 1 (§1.4.4), Rawls asserts that nonideal theory depends on ideal theory. He does not, however, clarify and defend this relation of dependence in any depth. I have clarified and defended the transitional nature of this relationship using a two-part argument. Given that nonideal theory is transitional it presupposes the content of ideal theory: the realization of ideal theory is the ultimate goal in each and every set of nonideal conditions. Ideal justice is uniquely privileged because it is the only stage that is relevant to
determining what we should do to remedy injustice at each and every stage prior to the realization of the ideal.

The realization of ideal justice is not, however, the only thing that is relevant at each and every stage prior to the realization of ideal justice. This is because *sui generis* principles of nonideal justice guide this transition. This leads to an important refinement of Rawls's position. As I noted in chapter 1 (§1.4.4), Simmons following Rawls argues that ‘good policies are good not relative to the elimination of any particular, targeted injustices, but only relative to the integrated goal of eliminating all injustice.’ However, this is false. Nonideal policies are not only good relative to the integrated goal of eliminating all justice (i.e., ideal theory). They are also good and bad relative to the nonideal principles of justice.\footnote{One reason this refinement is important is that it overcomes a potential objection to Rawls’s position as it stands. Hye Ryoung Kang contends that the problem with the claim that policies are good only relative to the integrated goal of eliminating all justice is that ‘…the policy “to take one step backward in order to take two steps forward” can be justified only when these steps are applied to the same group of people at a time. If the burdens from taking one step backward are assigned to one group of people, and the benefits from taking two steps forward to another group of people, whether these are future or present people, then this policy is unfair. If omissions of this kind are systematically applied and contribute to one specific group’s advantage or to another specific group’s disadvantage, such a theory of justice may be even more seriously flawed, because the theory loses its basis in impartiality, and as a result justice comes to be regarded as the advantage of the majority or of the strong, in contrast to Rawls’s intention.’ (Hye Ryoung Kang, “Can Rawls’s Nonideal Theory Save his Ideal Theory?,” *Social Theory and Practice* 42 (2016): 32-56, p. 53.) The introduction of *sui generis* principles of justice prohibits treating people in this problematic way.}

Of course, the set of nonideal principles of justice is not sufficient, in and of itself, to establish that nonideal theory presupposes ideal theory. This is because even if such nonideal principles of justice presuppose the content of ideal theory, some additional argument is required to show that nonideal principles of justice are necessary in order to do nonideal theory. It could, for instance, be argued that we can reliably judge what should be done in nonideal conditions without antecedently determining nonideal principles of justice. In chapter 3, I clarify the sense in which such principles are necessary.
2.9 Generalizing the results of my analysis

As I stated above, in determining nonideal principles of justice, the contracting parties assume that Rawls’s ideal theory of justice is correct. However, suppose that this assumption is false. If this is the case then what Rawls says in the following passage is pertinent:

\[
\text{[J]ustice as fairness, like other contract views, consists of two parts: (1) an interpretation of the initial situation and of the problem of choice posed there, and (2) a set of principles which, it is argued, would be agreed to. One may accept the first part of the theory (or some variant thereof), but not the other, and conversely.}^{55}
\]

Suppose, therefore, that the contractualist procedure that Rawls uses to determine ideal principles of justice and the analogous procedure that I have used to determine nonideal principles of justice are accepted, but it is argued that the output of the ideal contractualist procedure would be different. For instance, it is concluded that a mixed conception would be adopted that substitutes Rawls’s conception of the difference principle with a principle of average utility subject to the constraint that a certain social minimum is maintained.\(^{56}\) If this is the case then the nonideal principles of justice will have a different content, however, my preceding analysis will still be helpful because it reveals, at least if we accept the ideal and nonideal contractualist frameworks, that there are certain ways in which nonideal theory must relate to ideal theory. Clauses II-IV will be identical because they make no reference to the specific content of ideal theory. More interestingly, the following seems to hold whatever the content of both ideal and nonideal principles of justice happen to be:

Relationship between nonideal theory and ideal justice: In nonideal conditions we should transition fairly to the realization of ideal justice. (What constitutes a fair transition is specified by the principles of nonideal justice).

---


\(^{56}\) See: Ibid. 107.
The only exception is if the ideal contracting parties were to select a single teleological principle such as act utilitarianism. If this were the case we could not accurately say that nonideal theory involved transitioning to an end goal subject to certain constraints: a single principle “maximize utility” would apply under all conditions whether nonideal or ideal.

As I acknowledge above, the nonideal principles of justice also presuppose the economic conditions of relative affluence. Interestingly, towards the end of *A Theory of Justice* Rawls himself speculates about the content of justice if such economic conditions do not obtain. He writes:

I have supposed that if the persons in the original position know that their basic liberties can be *effectively* exercised, they will not exchange a lesser liberty for greater economic advantage (§26). It is only when social conditions do not allow the full establishment of these rights that one can acknowledge their restriction. The equal liberties can be denied only when it is necessary to change the quality of civilization so that in due course everyone can enjoy these freedoms.57

Under such economic conditions what Rawls terms the “general” as opposed to the “special” conception of justice—ordered by priority rules—applies. As Korsgaard explains the general conceptions ‘…tells us that all goods distributed by society, including liberty and opportunity, are to be distributed equally unless an unequal distribution is to the advantage of everyone, and especially those who fall on the low side of the inequality.’58 The special conception removes ‘…liberty and opportunity from the scope of this principle and says they must be distributed equally, forbidding tradeoffs of these goods for economic gains.’59

Rawls’s suggestion seems plausible. If the contracting parties knew that society was in such impoverished conditions that it had insufficient resources to allow its citizens to effectively exercise their basic liberties, they would not insist that the protection of such liberties was lexically prior to other economic goods. This is because such insistence would

57 Ibid. 474-475 (emphasis added).
58 Korsgaard, *Creating the Kingdom of Ends*, 147.
59 Ibid.
not maximize their interests: it would be pointless to insist on something that they knew could not be achieved.\(^6\) Furthermore, the claim that endorsing this general conception of justice under such conditions would lead to a methodological tension with Rawls’s serially ordered ideal principles of justice seems implausible: under such conditions there is a principled reason for relaxing the constraint of priority rules because such priority rules cannot be effectively realized under such conditions.

However, despite the fact that Rawls’s remarks about the general conception of justice seem plausible, two points are worth emphasizing. First, Rawls’s theory of justice is incomplete: the special and general conception are not sufficient to fully regulate all empirical conditions; under relatively affluent nonideal conditions the set of nonideal principles of justice that I have identified apply. Second, the derivation of the set of nonideal principles brings into focus how different principles required in conditions of unfortunate non-compliance due to poverty are from principles required in nonideal conditions that have sufficient economic resources. In the latter set of conditions—which obtain in actual countries like the present day United States—the demands of justice are still structured according to priority rules.

\(^6\) Within Rawls’s work the level of what economic development is sufficient for this general conception not to apply is distressingly vague. Robert Taylor has attempted to sharpen this component of Rawls’s theory. He distinguishes between the following three interpretations of the level: \textit{Formal Threshold}: Before the Priority of Liberty can apply, a society must have achieved a level of wealth sufficient for it to maintain a legal system with courts, police, and so on, that can define and protect the basic liberties of citizens; \textit{Weak-Substantive Threshold}: Before the Priority of Liberty can apply, a society must have achieved a level of wealth sufficient for it to allow its citizens to engage in the meaningful \textit{formation} of life plans. For example, citizens must have access to public forums, and have sufficient leisure time to make use of these resources and reflect on their plans; \textit{Strong Substantive Threshold}: Before the Priority of Liberty can apply, a society must have achieved a level of wealth sufficient for it to allow its citizens to engage in meaningful \textit{advancement} of life plans. For example, citizens must have access to professional training, start-up funds etc. He argues that the formal threshold is uncontroversially necessary because without it such liberties would be meaningless. However, ‘...the Strong Substantive Threshold must be ruled out. Once the Weak Substantive Threshold is met, our highest-order interest in rationality can be fully satisfied.’ (Robert S. Taylor, “Rawls’s Defense of the Priority of Liberty: A Kantian Reconstruction,” \textit{Philosophy & Public Affairs} (2003): 246-271, p 263.)
2.10 Conclusion

In this chapter I have derived and defended a set of nonideal principles of justice. In contrast to Rawls’s conception of nonideal theory that I surveyed in chapter 1, these principles give substantive normative content to nonideal theory, and clarify its relationship with ideal theory.
Chapter 3

THE VALUE OF NONIDEAL PRINCIPLES OF JUSTICE

3.1 Introduction

In chapter 2 I derived a set of nonideal principles of justice, and clarified the way in which they depend on ideal theory. This was an important initial step in a program to vindicate the value of ideal theory in actual nonideal conditions. This step was not, however, sufficient. It only answered what could be termed the “minor skeptical challenge”: it only established that there is a bridge between ideal principles of justice—that assume strict compliance and are determined in idealized counterfactual conditions—and nonideal principles of justice—that also assume strict compliance and are determined in idealized counterfactual conditions.

Establishing that there is a bridge between ideal and nonideal principles of justice in this sense is clearly insufficient to assuage the following major skeptical challenge: there is no way of bridging the gap between idealized principles of justice and the concrete questions that we actually face under nonideal conditions. How, for instance, do the nonideal principles of justice help determine what rate of progressive taxation is just under a specific set of conditions?

The outstanding task that therefore remains is giving an account of the role that the nonideal principles of justice should perform in the design and day-to-day operations of institutions in actual nonideal conditions. To clarify, I use the term “institution” in a very broad sense. It encompasses all the organizations which create, enforce, and apply policies and laws in the form of constitutional principles, federal policies, positive laws, legal precedents, etc.
This chapter has the following structure: I begin by showing that principles derived in idealized counterfactual conditions—such as the nonideal principles of justice—can at least in principle offer adequate normative guidance in actual nonideal conditions. I then argue that such principles perform a regulative function in actual nonideal condition. I defend this regulative function by responding to some arguments that attempt to establish that such principles are normatively and conceptually redundant. After that I argue that the instrumental value of justice in nonideal conditions cannot be reduced to this regulative function—it also performs an agent-guiding function, which directs agents to view and respond to the institutional realization of justice in the appropriate way.

3.2 Hye Ryoung Kang’s argument

Ideal theory has been extensively criticized for its failure to provide adequate guidance on actual forms of injustice (e.g., racial injustice). Kang notes one strategy that philosophers who are sympathetic to a broadly Rawlsian conception of the relationship between ideal and nonideal theory have used to respond to such criticism: to argue that they stem from a misconception of what ideal theory is designed to achieve, and an under-appreciation of the potential resources of nonideal theory. In particular, such criticism fails to recognize that the problems and limitations that arise because of the method of counterfactual idealization used to construct ideal theory can be addressed at the later stage of nonideal theory.

However, Kang’s central argument is that Rawls’s nonideal theory is unable to save his ideal theory. This is because principles of ideal justice that are derived under idealized counterfactual conditions are completely blind to certain forms of actual injustice (e.g., racial injustice). Consequently, nonideal theory of a transitional sort is unable to rectify the limita-

---

tions of Rawlsian ideal theory: if ideal theory is blind to certain actual injustices (e.g., racial injustice) and the exclusive function of nonideal theory is to transition to ideal justice it must also, in turn, be blind to certain actual injustices (e.g., racial injustice).  

My conception of nonideal theory is, in keeping with the Rawlsian tradition, transitional. However, it is different from that of other interpreters because it includes nonideal principles of justice. However, despite this difference Kang’s argument also applies to my position: if her argument succeeds then the innovation of nonideal principles of justice takes us no further in the quest to provide normative guidance under actual nonideal conditions. For, these principles are derived in idealized, counterfactual conditions and, therefore, must in principle be blind to certain types of actual injustice (e.g., racial injustice).

It is therefore imperative to begin by responding to her argument because if it is successful it would be devastating to my whole project: it could be determined a priori that the nonideal principles of justice were incapable of providing adequate normative guidance in actual nonideal conditions.

3.2.1 An outline of the argument

Kang argues that principles of justice that are derived under idealized counterfactual conditions are in principle flawed when they are applied to actual conditions because their application is an instance of what Robert Shope terms the “conditional fallacy.” Shope argues this fallacy is often committed in philosophical accounts that rely on counterfactual assumptions. The conditional fallacy has the following kind of subjunctive form:


5 Kang’s argument is the most rigorous and clear presentation of an objection that has bubbled under the surface of the literature on nonideal theory for quite some time. As Laura Valentini notes, the strongest version of the charge that has been presented against ideal theory is that at best it is morally irrelevant, at worst morally destructive. Valentini unpacks this line of objection by explaining: ‘if we apply principles developed under ideal conditions to real-world circumstances—namely those circumstances for which they have allegedly been designed—we are bound to obtain morally counterintuitive results.’ (Laura Valentini, “On the Apparent Paradox of Ideal Theory,” Journal of Political Philosophy 17 (2009): 332-355, p. 341.)

6 Kang begins her paper by claiming that the inadequacy of Rawls’s nonideal theory in The Law of Peoples—which addresses injustice in the context of globalization—is not accidental but inherent in Rawls’s method: the construction of ideal theory using counterfactual idealization. (Kang, “Can Rawls’s Nonideal
If state of affairs $a$ were to occur, then state of affairs $b$ would occur. A conditional fallacy is committed when, in a specified situation, if $a$ were to occur, then it would be at least a partial cause of something that would make $b$ fail to occur.\(^7\)

In order to establish that Rawls does in fact commit the fallacy, Kang reformulates the procedure that Rawls uses to derive principles of justice into the following simplified conditional form:

\[
\Phi \text{ [principles of justice] would be recommendable for } A \text{ [actual political actors] in } C \text{ [actual conditions] to regulate } S \text{ [actual political institutions] iff } A^+ \text{ [idealized agents] in } C^+ \text{ [idealized, counterfactual conditions] were to agree on choosing } \Phi \text{ to regulate } S^+ \text{ [ideal political institutions].}\(^8\)
\]

The consequent of the conditional, essentially, states that the demands that should apply under actual conditions are determined by what idealized agents in idealized, counterfactual conditions would assent to regulate ideal political institutions. Her reformulation is clearly an

---

Theory Save His Ideal Theory,” 44). However, towards the end of her article she anticipates the objection that her critique of idealization applies only to bad idealization and not idealization per se and, consequently, she crafts the argument that I discuss below, for why principles of justice derived under idealized counterfactual conditions are in principle inadequate for the task of offering normative guidance in the actual world. I think that it is important to note that the problems that Kang uncovers with Rawls’s theorizing in The Law of Peoples cannot be an illustration of what is in principle wrong with the method of using principles of justice derived under idealized conditions. From the premise that there are substantive problem with the way in which an ideal model is constructed in a particular instance (e.g., in The Law of Peoples, where Rawls unsatisfactorily models the states as autonomous and self-sufficient), we cannot validly infer the conclusion that there is something problematic with relying on the method of idealization per se. I think that it is absolutely crucial to distinguish the question of whether the method of counterfactual idealization is in principle unable to offer normative guidance in actual nonideal conditions from the question of whether a particular theorist in a particular instance uses this method to construct an unsatisfactory model of ideal and nonideal justice.


\(^8\) Kang, “Can Rawls’s Nonideal Theory Save his Ideal Theory?” 49.
accurate representation of the contractualist procedure that is used both by Rawls to determine ideal principles of justice, and the analogous contractualist procedure that I have used to determine nonideal principles of justice.

The main argument that Kang advances for the conclusion that such a procedure is in fact an instance of the condition fallacy is the following:

[S]uppose that ACS includes circumstances of justice historically and socially situated, such as the current realities of racial or gender justice that do not exist in ACS+. Given the lack of those circumstances, it would not be theoretically possible for ACS+ to choose principles (π) or resources to address racial or gender problems that would not exist in the counterfactual circumstances. By definition of A+, it would not make sense for A+ to choose principles of justice for those circumstances that do not exist in ACS+, because choosing such principles would not be reasonable for A+… [therefore] Φ that is chosen by ACS+ is necessarily misguided for ACS.⁹

Although Kang’s argument appears to be technical because it contains so many variables, in essence, the claim that is being made is simple and intuitive: the idealized counterfactual conditions under which principles of justice are determined are radically different from actual nonideal conditions. In particular, under such counterfactual conditions, in which agents are trying to maximize their individual prospects, considerations such as racial injustice do not arise because racial injustice is something that is idealized away under these conditions. Therefore, principles determined within such counterfactual conditions cannot in principle offer guidance in the actual world on forms of injustice such as racial injustice because such injustice is not present in the idealized conditions. The Rawlsian approach is an instance of the conditional fallacy because the principles of justice that would be selected under the idealized counterfactual conditions are guaranteed to be different from the considerations of justice that apply in the actual case, although they are falsely assumed to be identical.

---

⁹ Ibid. 49-50.
3.2.2 Response to Kang’s argument

Kang’s argument hinges on a descriptive claim: the conditions in an idealized counterfactual procedure are different from the conditions in the actual nonideal world; whilst subsets of actual people are victims of, for instance, racial injustice, the contracting parties are ex hypothesis not victims of racial injustice because the facts that constitute racial injustice are idealized away in the contractual procedure. Kang is clearly correct about this descriptive distinction. However, what is problematic with her argument is the conclusion that she infers from this descriptive distinction: because the conditions are different in this descriptive sense she concludes that it is in principle impossible to generate norms under an idealized counterfactual procedure that offer adequate normative guidance for conditions that contain types of injustice not present in the idealized counterfactual procedure. This conclusion is false. It presupposes the false tacit premise that norms that are generated under idealized counterfactual conditions cannot, in any way, offer normative guidance under conditions that are different from the contractualist procedure. Or, in other words, that in order to offer normative guidance on a feature of the actual world X, X must also be a feature of the contractualist procedure.

The example of racial injustice is the best way of illustrating problems with Kang’s argument. Under idealized counterfactual conditions the parties select a first principle of ideal justice, which (roughly) states that each person has an equal right to the most extensive scheme of basic liberties. This principle makes no explicit reference to racial injustice. However, it does have substantive implications for racial injustice. It straightforwardly condemns, for instance, race based slavery. For, under such an institutional arrangement not everyone has access to the most extensive scheme of basic liberties.

3.2.3 Final Comment

Of course, Kang’s challenge could be reformulated as a weaker argument: even if it is not in principle impossible for an idealized counterfactual procedure to provide adequate normative guidance in the actual would it seems unlikely that it will offer the correct type of guidance. This is because the conditions in which the contracting parties are placed are so different from the conditions in which we find ourselves in the actual world. However, all I
want to establish at this stage of the argument’s dialectic is that Kang’s claim that principles of justice derived under idealized conditions are in principle incapable of providing guidance in actual nonideal conditions cannot be sustained. Of course, whether nonideal principles of justice are in fact valuable is an outstanding question.

3.3. The regulative function of the nonideal principles of justice

I argue that the nonideal principles of justice play a regulative function in the design and operations of nonideal political institution. By this I mean that they guide the judgments of political actors. This notion of “offering guidance” must be distinguished from that of “trying to directly or fully realize.” This latter concept is the wrong conception of the nonideal principles of justice’s application to actual political practice: principles of justice that presuppose actual people’s strict compliance and fail to take into account considerations of political feasibility cannot, plausibly, be directly applied to conditions in which these variables are present. They are, rather, principles that should offer guidance mediated by good judgment.

3.4 Defending the regulative function of the nonideal principles of justice

I will defend the regulative function of these nonideal principles by responding to three objections. I will use the same strategy to diffuse each objection: I will show that they all depend (in one way or another) on an implausible conception of how such principles

---

10 The term “regulative” is used in different senses by various philosophers. My use of the term differs from that of John Searle and Kant. Searle construes a rule as regulative if it regulates pre-existing activity whose existence is independent of the rules. (See: John R. Searle, The Construction of Social Reality (USA: Free Press, 1997).) In the Transcendental Analytic of the First Critique Kant argues that regulative principles (e.g., the analogies of experience) apply to relations between existences (A179/B222). (For further discussion see Gary Banham, “Regulative Principles and Regulative Ideas,” in Kant Unde Die Philosophie in Welburgerlicher Absicht: Akten des Xi. Kant-Kongresses, ed. Margit Ruffing et al. (2010) 15-24.) Kant’s use is at least roughly comparable to that of Searle in that these principles “regulate” pre-existing activity; these regulative principles are not like the constitutive ones (e.g., the axioms of intuition) that determine the “quantification” of experience itself (i.e., the very possibility of the existence of discrete intuitions of the world). I think that it is fair to say that my use of the term is closer to at least one ordinary language use of the term as “something that guides.”
should apply to actual political practice. Once these implausible conceptions are exposed the objections dissolve, and the function and value of the nonideal principles of justice is further clarified and defended.

The objections that I will consider were originally presented against ideal theory. However, they also apply mutatis mutandis to the nonideal principles of justice that I have derived and defended. For, these nonideal principles of justice assume strict compliance and are determined in idealized counterfactual conditions and, therefore, are also “ideal” in the requisite sense for the same objections to apply.

3.4.1 The indeterminacy concern

As I acknowledged above, the relationship between the nonideal principles of justice and the actual operations of political institutions is not one of deductive entailment. However, this gives rise to the objection that the guidance provided by these principles is indeterminate. Norman Daniels’s work illustrates this concern. Throughout his career, Daniels has tried to develop a Rawlsian account of access to health care. In his early work he argued that a robust notion of Rawlsian equal opportunity is sufficient to generate and justify a set of social policies bearing on access to health care. However, in his late work he came to acknowledge that philosophical theory is not, by itself, sufficiently fine-grained for such concrete policy-making and must, therefore, be supplemented by something like political deliberation.11

I acknowledge that there is a degree of indeterminacy in the application of principles to practice. However, this degree of indeterminacy is unproblematic. As the prefatory epigram of my dissertation conveys: all principles require interpretation (in the form of good and reasonable judgment) when they are applied to actual cases.12 It is this ineliminable role

---


of good judgment that accounts for the indeterminacy of theory in independence from such judgment.

The best way of defending the claim that a degree of indeterminacy is not a good reason to reject the value of nonideal principles of justice is to reflect on what would be the case if alternative, possible approaches were adopted. Such alternative possible approaches are either deeply implausible or else also require conceding that there is a degree of indeterminacy.

In the early editions of Tom Beauchamp and James Childress’s seminal work *The Principles of Biomedical Ethics* they included a flow chart moving from moral theory, to principles, to rules and finally to case judgments.\(^{13}\) Such a model at least engendered the impression that the flow from principles to actual cases was immediate and determinate. However, as John Arras notes, Beauchamp and Childress ‘…have long since explicitly repudiated both that diagram and the account of moral principles implied by it.’\(^{14}\) Such a conception of how principles apply to actual cases has the theoretical allure of scientific rigor and precision. However, it is not a conception that can sustain critical scrutiny. Any moral principles—except a single-valued consequentialist principle—will require judgment when it is applied to actual cases.

Alternatively, if like Amartya Sen a philosopher rejects *tout court* the value of idealized principles of justice in actual nonideal cases, they will still have to ultimately concede that such an approach will involve a comparable degree of indeterminacy. Indeed as Gilabert persuasively notes:

> *Sen is surely right than an ideal theory [and by extension nonideal principles of justice] would not provide any algorithm for settling tradeoffs when comparing all feasible alternatives that fall short of the ideal. But neither does Sen when he address the issues in the book [The Idea of Justice]. He should not*

---


fault ideal theorizing for not delivering what it does not try to yield, and what
his own preferred approach does not deliver.\(^{15}\)

Ultimately, an adequate defense of the value of nonideal principles of justice in actual
nonideal conditions should not require the bar to be set absurdly high. The value of such
principles can be affirmed whilst acknowledging that a degree of indeterminacy is an ines-
capable feature of the mediate application of these principles to actual nonideal conditions.

3.4.2 The appeal to ideal and nonideal principles of justice is conceptually redundant

Michael Philips argues that principles of justice are conceptually redundant in actual
nonideal conditions. He writes:

Each of these ideal moralities consist of a coherent set of principles that—
taken together—promotes some goal (e.g., justice for Rawls, unalienated
humanity for Marx). This being so, perhaps we can adapt these moralities to
a set of historical conditions by asking what principles would promote these
goals under these conditions. The problem is, however, that if this is our tac-
tic, we are no longer deriving anything from the Ideal Principles. The Ideal is
idle; our derivation is directly from the goal itself.\(^ {16}\)

Of course, on my conception of nonideal theory, there is not an outstanding question of
how we get from ideal principles of justice that specify how a perfectly just society should be
ordered to principles that regulate not-entirely-just institutions: this is explained by the
nonideal principles of justice. However, the question still remains of what essential role
“principles” of justice are performing.

The answer, I believe is that (at least on a broadly Rawlsian approach) we cannot talk
of a goal of justice in independence from principles of justice; this is because for a Rawlsian

\(^{15}\) Pablo Gilabert, “Comparative Assessments of Justice, Political Feasibility, and Ideal Theory,” *Ethical

\(^{16}\) Michael Phillips, “Reflections on the Transition from Ideal to Non-Ideal Theory,” *Noûs* 19 (1985):
551-570, p. 563.
principles of justice are constitutive of justice. This response, I concede, might seem unsatisfactory because it may seem as though I am responding to Philips by simply invoking a stipulation about the conceptual nature of justice. However, even if justice were construed as a goal that should be realized (e.g., as a state of affairs that realizes certain properties such as equality) the demands of nonideal justice cannot be captured in terms of the realization of such a goal. This would collapse nonideal justice into the teleological conception that I considered and rejected in chapter 2 (§2.7.1): justice is not just a goal that should be realized in the future; there are restrictions on how this goal can be realized. Therefore, principles of ideal justice and further nonideal principles of justice derived from them are indispensable because they structure our responsibilities in actual nonideal conditions.

3.4.3 The appeal to ideal and nonideal principles of justice is normatively redundant

As I noted in chapter 1, the orthodox Rawlsian position is that ideal theory stands in a priority relation to nonideal theory. Sen famously challenges this claim: in order to determine what should be done in nonideal conditions, it is not necessary to antecedently determine an ideal theory of justice. Note that my previous arguments are not sufficient to overcome this objection: showing that a set of nonideal principles of justice depend on ideal theory, and that these nonideal principles of justice can perform a regulative role in actual nonideal conditions is not sufficient to show that they must perform such a regulative role.

In his work on health care and international justice, Gopal Sreenivasan offers one of the most interesting elaborations of Sen’s line of argument:

[T]here exists a kind of non-ideal theory for which the priority assumption fails. On this conception, non-ideal theory functions as an anticipation of ideal theory. Its prescriptions anticipate the ideal requirements of justice rather than presupposing them. To do so, non-ideal theory has to make assumptions about the minimum requirements that any plausible and complete

---

ideal theory of justice will include. In this vein, it can define targets for practical action before a complete ideal has been worked out, even in outline.\textsuperscript{18}

As Sreenivasan goes on to explain, he has in mind the following example of nonideal theorizing that can “anticipate” the results of ideal theorizing:

\ldots any plausible and complete ideal theory of international distributive justice will minimally include an obligation on the richest nations to transfer 1 percent of their GDP to the poorest nations.\textsuperscript{19}

In order to assess this objection to the claim that ideal theory stands in a priority relation to nonideal theory, it is crucial to distinguish between the following two relations of dependence:

A guidance relation of dependence: ideal (and by extension nonideal) principles of justice play an essential role in guiding the actions of political actors; consequently, without such principles the actors would not be able to determine what should be done.

A theoretical relation of dependence: the best theoretical account of what should be done in nonideal conditions requires appealing to ideal (and by extension nonideal) principles of justice, regardless of whether such principles play an essential role in guiding the actions of political actors.

Note that Sen and Sreenivasan have only produced arguments against the former relation of dependence. Even if the guidance relation of dependence fails to hold this does not entail that the theoretical relation of dependence does not hold. Indeed, I suggest that the theoretical relation of dependence always holds: if political actors are confronted by abhorrent injustice (e.g., race-based slavery) they may well be able to reliably judge what should be


\textsuperscript{19} Ibid.
done without the guidance of ideal and nonideal theory. However, such reliable judgment should not be conflated with a theoretical account of why they should act in this way. Even in such a clean-cut case, I think that the most plausible theoretical account is that they are acting to fairly transition to a more just world, as specified by ideal and nonideal principles of justice.

However, Sen and Sreenivasan are clearly correct that the guidance relation of dependence does not hold in all cases. An array of counter-examples can be produced against the claim that ideal and nonideal principles always perform an essential role in guiding the actions of political actors.

It is worth noting—as Sen and Sreenivasan do not—that this conclusion should not be all that surprising. It is, in fact, entailed by a commitment to a reflective-equilibrium methodology. As Rawls himself explains the purpose of a contractualist procedure is to model and extend our intuitions about justice, which we take as provisional fixed points. Consequently, of course it is possible—in an array of cases—to determine what should be done in independence from principles of ideal and nonideal justice. After all, such principles are, themselves, constructed (at least partially) by obdurate intuitions about an array of actual cases.

I think that the best response to this line of objection is to argue that the guidance relation of dependence cannot be as crude as Sen and Sreenivasan implicitly presuppose. Of course, there are a range of cases for which the guidance relation of dependence fails to hold. However, nonideal theory does not just consist of clear-cut cases like slavery and minimal provisions for global health. There are a number of far more contested topics. For instance, what are the conditions, if any, under which non-meritocratic affirmative action is just? As I will tentatively argue in chapter 5, for this type of case the guidance supplied by ideal and nonideal principles of justice may well be indispensible.

Interestingly, I think that a careful reading of Rawls reveals that he is sensitive to the very concerns that Sen and Sreenivasan present as objections against his theory. Towards the beginning of A Theory of Justice Rawls writes that he begins with ideal theory because he believes that it ‘…provides, I believe, the only basis for the systematic grasp of these more

---

pressing problems…\textsuperscript{21} of nonideal theory. Here the caveat of “systematic” is important: he is not suggesting that it is essential in all cases, just that it is necessary to obtain a systematic grasp of the subject matter of nonideal theory. This claim seems plausible to me and the individual counter-examples produced by Sen and Sreenivasan are not sufficient to refute it.

3.5 Introducing the agent-guiding function of the nonideal principles of justice

I have argued an important instrumental value of the nonideal principles of justice is its institution-guiding regulative function. However, I also want to argue that the instrumental value of the nonideal principles cannot be reduced to this instrumental function. They also perform an agent-guiding function, which facilitates the right response to the institutional realization of justice. This agent-guiding function of justice has been overlooked. Bringing it into focus is crucial to justifying the methodological assumption of strict compliance used to determine the ideal and nonideal principles of justice: agents need a conception of justice that is not tainted by the non-compliance of actual agents, in order to respond to feasible institutional realizations of justice that are appropriately attuned to actual agents’ expected non-compliance.

3.5.1 How I will uncover this agent-guiding function

I will uncover this agent-guiding function by exploring the following epistemic question: how, if at all, does the value of a conception of justice depend on the epistemic limitations of actual agents? Consider the familiar example of Rawls’s Difference Principle.\textsuperscript{22} The institutional realization of that principle would require agents to determine which policies (e.g. rates of taxation over the next ten years) would maximize the expectations of the least advantaged. This, clearly, would require them to make incredibly difficult long-term predictions.

It is worth pausing to acknowledge that my nonideal theory of justice imposes far weightier epistemic demands on agents than the sketch of nonideal theory that we receive

\textsuperscript{21} Ibid. 8. (emphasis added).

\textsuperscript{22} For the fullest statement of the difference principle see: John Rawls, \textit{A Theory of Justice: Revised Edition} (Cambridge Mass.: Belknap Press, 1999), 266.
from Rawls. Clearly it is going to be hard enough to compile enough social scientific data to insure that we are roughly transitioning, subject to the threshold constraint of “moral permissibility”, to the realization of ideal justice; it will be far harder still, under many sets of nonideal conditions, to know that we are transitioning to ideal justice via a perfectly just transitional path as specified by the nonideal principles of justice. Assuming that my defense of the nonideal principles of justice has been successful, the concern might arise that the “normative deliverance” of Rawlsian nonideal theory has come at very hefty epistemic expense.

The concern underlying this epistemic question can be capture more precisely: is the following condition necessary for a conception of justice to be valuable:

Epistemic Access (EA): a conception of justice is instrumentally valuable in nonideal conditions if and only if actual agents have epistemic access\(^{23}\) (in some way to be further specified) to the particular requirements for realizing this conception of justice.

This condition is *prima facie* plausible: it seems as though such epistemic access is necessary for a conception of justice to be useful. However, as I shall show, we should reject EA. My novel argument runs as follows: in order for agents to adopt the right responses to institutional realizations of justice that reflect their epistemic limitations, they must have a conception of justice that is graspable independent of those same limitations. Essentially, a conception of justice does not merely guide institutional realizations of justice; it also guides agents to view and respond to these institutional realizations in the appropriate way.

---

\(^{23}\) This concept of “epistemic accessibility” can be parsed in different ways. For instance, it could be argued that the extent to which the relevant beliefs (about what is required to realize this conception of justice perfectly) are warranted must reach a certain justificatory threshold. Or, to adopt a weaker variant of essentially the same position, the relevant beliefs of a subset of actual agents—perhaps those agents placed in politically key positions—must reach that threshold. Alternatively, it might be urged that even if agents do not need to know enough to be confident that they are realizing the demands of justice perfectly, they do need to know enough to be sure that they are realizing something at least in the “ball park” of how one would go about realizing the demands of justice. The explication of the concept will also depend on whether an internalist or externalist conception of justification is adopted. For my present purposes, these complexities can be sidelined: regardless of which way the concept of epistemic accessibility is construed my central argument holds.
My argument has the following structure: I begin by outlining the intuitive case both for and against EA. After that, I consider a *prima facie* plausible but ultimately unsuccessful argument against EA, in order to illuminate what a successful argument against EA must establish. I then present a two-part argument against EA, and show that an analogous argument can be constructed for political feasibility. Finally, I trace out the implications of my analysis for prominent proponents of nonideal methodology (i.e., a methodology that rejects ideal theory).

3.5.2 Some initial clarifications

I present EA with respect to “a conception of justice” rather than “an ideal theory of justice” in order to make clear what the various arguments are and are not sufficient to establish. In order to defend the value of an ideal theory of justice in nonideal conditions—in which the realization of this ideal is not epistemically accessible—it is necessary to reject EA. However, a rejection of EA is not sufficient to defend the value of an ideal theory of justice. This is because it is possible to argue that the value of a conception of justice can outstrip the epistemic capabilities of actual agents without thereby endorsing the value of an ideal (i.e., perfect) theory of justice.

An agent-guiding function directs agents to adopt the appropriate responses to the institutional realizations of justice, in the form of both attitudes and actions. This agent-guiding function is distinct from the institution-guiding function in that it regulates the attitudes and actions of agents in relation to institutions rather than the design and operation of the institutions themselves.

3.6 Conflicting intuitions about EA

It is reasonable to have conflicting intuitions about EA. On the one hand, begin by considering the intuitive case for rejecting EA. Suppose that a philosopher argues against, for instance, act utilitarianism by showing that we cannot know which institutional arrangements will most effectively maximize net-utility. Her argument does not seem nearly as strong as the argument that by condoning abhorrent acts such as the punishment of inno-
cent people in actual or counterfactual scenarios, act utilitarianism supplies the wrong normative standard of justice.

One plausible diagnosis of this contrast in argumentative power is that with respect to the latter type of disagreement it is hard to envisage how such a critic could engage in a constructive discussion with someone who defends act utilitarianism. In contrast, one philosopher’s epistemic objection to another’s conception of justice could still leave enough common ground for the two to engage in further valuable debate.

There is, therefore, reason to think that epistemic objections to a conception of justice are weaker than objections against the normative standard enshrined in that conception of justice. (Of course, the precise theoretical upshot of this claim must be carefully considered: we must determine, in particular, whether it motivates the rejection of EA, or whether it supports the conclusion that epistemic objections are weaker but still sufficient to refute a candidate conception of justice at least under certain parameters.)

On the other hand, we might think that EA should be endorsed. Consider the following argument:

P1: If the requirements for realizing a conception of justice are epistemically inaccessible to actual agents (at least to a sufficient threshold) then this conception of justice cannot guide the design and operation of institutions.

P2: The purpose of a conception of justice is to guide the design and operation of institutions; theories that fail to fulfill this purpose should be rejected.

C: If the requirements for realizing a conception of justice are epistemically inaccessible to actual agents (at least to a sufficient threshold), the conception of justice should be rejected.24

---

24 An argument of this sort is presented by Liza Herzog: she argues that if we understand ideal theory as action-guiding then the unavailability of knowledge at the level of nonideal theory can lead to “feedback loops” between different levels of theorizing. This is because an ideal theory should be revised so that it is in line with the epistemic limitations of agents that are revealed at the level of nonideal theorizing. (Lisa Herzog, “Ideal and Non-Ideal Theory and the Problem of Knowledge,” Journal of Applied Philosophy 29 (2012): 271-288.)
If this argument succeeds it leads to exactly the same conclusion as an argument that establishes that a candidate conception of justice enshrines the wrong normative standard: the candidate conception of justice should be rejected. All that differs are the reasons that the conclusion is reached.

In summary, there is an intuitive basis for rejecting EA. However, this is not sufficient to justify rejecting EA. There is also a prima facie plausible argument for endorsing EA.

3.7 An unsuccessful argument against EA

Given that an intuitive case can be presented both for and against EA, we must turn to more elaborate philosophical arguments to try to clarify our intuitions and to iron out any inconsistencies. I begin by considering an argument against EA that is ultimately unsuccessful. Examination of this argument will, however, help to illuminate what type of argument must be made in order to justify rejecting EA.

The argument begins with a traditional conception of the division of labor between political philosophy and social science: it is the task of political philosophy to articulate “pure,” context-independent normative prescriptions (e.g., an ideal theory of justice). In contrast, as Adam Swift writes, it is the task of the social sciences ‘…to (try to) tell us what states of the world can indeed be realized by what means—with what probabilities, over what time scales—given where we are now.’ Therefore, epistemic considerations are out of place in the subject matter of “pure” political philosophy. They should be introduced only when such context-independent prescriptions are applied.

However, this argument is clearly inadequate as it stands: mere appeal to this traditional division of labor does not justify rejecting EA. This is because the traditional division may be no more than an unjustified dogma of political theorizing. We must, therefore, supply proper justification for this division.

---


One argument in defense of this division rests on the following feature of our epistemic predicament: under nonideal conditions, our ability to form predictions about the future is not fixed. Economists, for instance, can build ever more sophisticated models to forecast future levels of economic wellbeing. Given that our epistemic abilities are continuously evolving, we should not factor epistemic considerations into our most abstract conception of justice. This abstract conception should exclusively enshrine what we judge to be right, regardless of our epistemic limitations; epistemic considerations should be factored in only when we apply this conception of justice in an institutional policy-and-law-making context. Therefore, the argument concludes, EA should be rejected because epistemic accessibility is not necessary.

This argument fails because its conclusion is a non sequitur. Even granting the premise that we should distinguish between purely normative and epistemic considerations, because the latter are constantly shifting, the conclusion that our most abstract conception of justice should exclusively reflect purely normative considerations does not follow. For we could distinguish epistemic and purely normative considerations and still think both must be factored into the most abstract theorizing about justice. The argument rests on a tacit, undefended assumption. Either the argument assumes that our most abstract conception of justice should encode purely normative considerations because these are somehow more fundamental, or the argument assumes that it is theoretically desirable to have a conception of justice that is fixed and not dependent on our shifting epistemic capacities. Whichever its assumption, the argument begs the question against endorsing EA as presented above: the argument assumes, without supplying a principled reason, that P2 is false and that our most abstract conception of justice should not be continually revised in line with the shifting epistemic capacities of actual agents such that it can guide the institutional realization of justice.

The lesson that should be learnt from this unsuccessful argument is that the rejection of EA cannot be justified merely by stipulating the domain of political philosophy, or by appealing to the fact that the epistemic capacities of agents are ever shifting. Rather, a

---

successful argument against EA must explain what value a conception of justice that exceeds the epistemic capabilities of actual agents could possess. This is because, if it can be established that it is possible for the value of justice to outstrip the epistemic capabilities of agents, then it cannot be stipulated that it is necessary for the realization of a conception of justice to be epistemically accessible in order to deem that conception valuable.

An adequate assessment of EA, therefore, requires us to take a position on the value and purpose of political theorizing itself: should an (instrumentally valuable) political philosophy \textit{just} regulate the design and operation of institutions—using principles that make epistemically accessible demands—or should it also perform an additional function that cannot be reduced to this institution-guiding function?\textsuperscript{28}

3.8 A two-part argument against EA

An understandable pessimism about our ability to engage critically with EA may arise: it hinges on such a deep disagreement about the value and purpose of political theorizing that it is not possible to craft any intellectually persuasive arguments. I grant that the disagreement is very deep. I maintain, however, that there is a (relatively) neutral position from which EA can and should be assessed: its ability to capture the phenomenology of what agents value in nonideal conditions.\textsuperscript{29} In what follows I present a two-part argument

\textsuperscript{28} A different way of putting this point is to say that we must determine how to explicate the familiar dictum “ought implies can.” Does an endorsement of this dictum mean that the implications of our most fundamental normative standards (e.g., an ideal theory of justice) must be epistemically transparent to actual agents? Or is the dictum, as Robert Stern argues, not something that can be defended as a principle that limits what is right and wrong in the most fundamental sense, but only as a limit on the ability of actual agents to act on principles of rightness and to be assigned blame? (See: Robert Stern, “Does ‘Ought’ Imply ‘Can’? And Did Kant Think It Does?, Utilitas 16 (2004): 42-61.)

\textsuperscript{29} To qualify: my description of “actual agents” is slightly idealized. This is because there may be some actual agents who do not have political values of the sort that I will specify (or indeed any political values!). However, it would be unfair to see this as a counterexample to my position: my claim is that actual agents (who are informed and responsive to political values in the appropriate way) will have values that are not reducible to principles that guide the institutional realization of justice, given their epistemic limitations. Therefore, if the reader prefers, my position can be recast as a thesis about what actual agents \textit{should} value in nonideal conditions.
that shows that the phenomenology of what actual agents value about justice can outstrip their epistemic capabilities.

3.8.1 Two cases

Case 1: A young hiker named Likitha plans to hike a stretch of the Appalachian Trail in North Carolina. Before she sets off, she plans to make a detour between Standing Indian Mountain and Mount Albert to climb a steep peak in order to view a beautiful vista. However, when the day of the envisioned detour arrives, the potential perils of embarking on it dawn on her: she does not know the terrain and the quality of the signposting is deeply uncertain. She prudently decides not to undertake the detour.

Case 2: A group of anthropologists are studying the social infrastructure of a recently contacted tribe. In their initial investigations they only uncover a conception of justice that is fully attuned to the tribe’s epistemic limitations. However, after a more thorough investigation, the anthropologists uncover another conception of justice that the tribe endorses—one that transcends the tribe’s epistemic capabilities.

I submit that Likitha’s plan reveals a dimension of value that is significant to her experience even though it is not, in the end, directly action-guiding. Likitha’s plan is significant not merely in the anodyne sense that she formed a particular plan that she subsequently had to revise. It also reveals a dimension of what she values about the hike: she values seeing beautiful vistas and judges *ceteris paribus* the viewing of a beautiful vista to be worth an investment of time and effort. Were the detour not to have possessed potential perils given her limited knowledge, Likitha would have embarked on the detour.

The anthropologists could view their more thorough investigations as unearthing a bizarre—and ultimately valueless—cultural practice that the members of the tribe have of forming idle fantasies beyond their epistemic capabilities. However, a more reasonable response would be to view the transcendent conception of justice as a cultural edifice that reveals something fundamental about the tribe’s values. Indeed, discovering such a concep-
tion seems to bring into focus a dimension of the tribe’s values that was previously obscure to the anthropologists. Just as the revelation that Likitha—absent certain epistemic limitations—would have valued going on a detour discloses something significant about her as an agent, so the revelation that the tribe—absent certain epistemic limitations—endorses a certain conception of justice reveals something fundamental about their deepest political values.

Considering the anthropologists’ investigations from a third-person perspective brings out the political values that we (at least tacitly) hold from a first-person perspective: under a wide range of conditions we do not only value a fully epistemically accessible conception of justice. Consequently, if EA were endorsed as a rigid constraint on a theoretical conception of justice, our theorizing would fail to capture an important dimension of value. Therefore, EA should be rejected.

3.8.2 A transcendental argument

Perhaps the intuition that I am trying to extract from these cases can be resisted: the anthropologists’ investigations are certainly intriguing, but there is no reason to think that the conception of justice that transcends the tribe’s epistemic limitations is valuable in some deeper sense. This is because, one might urge, despite the fact that the anthropologists label this transcendent conception as part of the tribe’s conception of justice it is not clear that it is appropriate for them to classify the transcendent conception as part of the tribe’s real conception of justice.

In response, I present the following transcendental argument:\(^{30}\)

In order for agents to adopt the right responses to institutional realizations of justice that reflect their epistemic limitations, it is necessary for them to have a conception of justice that is graspable independent of those same limitations.

---

30 Schematically transcendental arguments have the following form: X is a necessary condition for the possibility of Y. Y is the case. Therefore, X must also be the case. (Following Robert Stern, “Introduction,” in *Transcendental Arguments: Problems and Prospects* ed. Robert Stern (Oxford: Oxford University Press, 2003), 1-5.)
Consider the following illustration of this argument: for egalitarian reasons, a society introduces a policy designed to ameliorate the economic disadvantages of the least advantaged. This policy is appropriately sensitive to the policy makers’ epistemic limitations. However, the policy is ultimately unsuccessful because the policy makers are unable to successfully predict the effects of the policy, despite making a reasonable effort. The failure of this policy, which results from the policy makers’ epistemic limitations, can only be appropriately judged and responded to against the backdrop of a conception of justice that does not reflect those epistemic limitations.

This is so for two different reasons. First, it would not be reasonable for the least advantaged group to adopt the reactive attitude \(^{31}\) of blame toward the policy makers: after all, the policy makers acted for a good rather than bad reason. In order to adequately conceptualize a bad outcome ensuing from actions undertaken for good reasons—that is, to make sense of the bad outcome produced by the policy makers’ well-intentioned and reasonable actions—we must appeal to the following counterfactual: the policy makers would have implemented a conception of justice that transcends their epistemic limitations more effectively if they were able to do so.

Second, independent of their material wellbeing, it is important for the social standing of the least advantaged group that the community acknowledges (or at least ought to acknowledge) a conception of justice that transcends its epistemic limitations in a scenario in which a failure results because of such epistemic limitations. In the face of such a failure, it is still significant to the least advantaged group that people acknowledge that justice dictates that they were owed something—even if it was not, ultimately, possible to realize this demand of justice in a particular instance.\(^{32}\)

---


\(^{32}\) Following Stanley Cavell I use the term “acknowledge” rather than “recognize.” (Stanley Cavell, “Knowing and Acknowledging,” in *Must We Mean What We Say?* (Cambridge: Cambridge University Press, 1969), 238-266.) This is because, first, “recognition” can have loaded connotations of Hegelian metaphysics. (Following Mattias Iser, “Recognition”, *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta (2013): http://plato.stanford.edu/archives/fall2013/entries/recognition/.) Second, in contemporary ethical theory the concept is closely associated with Stephen Darwall’s work. Darwall argues that recognition is essentially second-personal (i.e., it is grounded in (de jure) authority relations that an addresser takes to hold between her and her addressee that cannot be reduced to a set of propositions). (Stephen Darwall, *The Second-Person*
In order to illustrate the profound importance of this acknowledgment given such a failure, consider a parallel scenario in which exactly the same outcome occurs (i.e., the material well-being of the least advantaged is not improved). However, in this parallel scenario failure does not result from the legislators’ epistemic limitations. Rather, it results from the fact that the legislators and the community at large regard the suffering of the least advantaged group as a matter of moral indifference. The material outcome of these two scenarios is identical. However, there is an important distinction between them: the original outcome results from an epistemic failure to realize a certain normative conception of justice, while the second does not result from an epistemic failure but rather from the rejection of this normative standard of justice. The intelligibility of this distinction depends, in part, on the fact that when a failure results because of epistemic limitations, it is salient that the community acknowledges a conception of justice that transcends these epistemic limitations.

The value of a conception of justice is not, therefore, reducible to an institution-guiding function that is fully attuned to the epistemic limitations of actual agents. It also has the value of enshrining the demands of justice in independence from the epistemic limitations of actual agents. This facilitates agents’ adoption of the right attitudes toward and actions in response to the demands of that conception of justice attuned to their epistemic limitations.

EA must, therefore, be rejected. Unlike the unsuccessful argument against EA that I examined above, I have now identified a principled reason to reject P2: the value of a conception of justice cannot be reduced to an institution-guiding function. Note: it is not merely possible, as the anthropologist example suggests, for the value of justice to transcend the epistemic limitations of the citizens that it governs (and therefore not necessary for the realization of the conception of justice to be fully epistemically accessible). Rather, if agents’ implementation of the demands of justice reflects their epistemic limitations and is fallible, it is necessary for them to have an agent-guiding conception of justice that transcends these limitations.

Standpoint: Morality, Respect, and Accountability (Cambridge Mass.: Harvard University Press, 2009.) I am neutral about whether Darwall’s thesis is correct. I simply argue (in a political context) that it is important for there to be norms that acknowledge people’s social standing, which do not reflect people’s epistemic limitations.
It is important to make two clarifications concerning the scope of this argument against EA. First, the argument does not presuppose that there cannot be certain epistemic constraints on a viable conception of justice. For instance, the argument allows that a candidate conception of justice could be rejected if it was sufficiently epistemically opaque and, consequently, could not perform either an institution-guiding or attitude-guiding function. The argument merely aims to show that in order for a conception of justice to be valuable it is not necessary for actual agents to have access to what would be required to achieve the full institutional realization of this conception of justice. Second, the argument does not presuppose that agents will always retrospectively be able to recognize the epistemic failure of previous attempts to institutionally realize justice. In some instances it will be fairly clean cut that an institutional arrangement has ultimately turned out to be disastrous. In other instances, determining whether an attempt to institutionally realize justice has been successful may be epistemically opaque, or at least be highly contested. For example, consider the example of the institutional realization of the Difference principle. This requires legislators to realize an economic distribution of goods that is to the greatest possible advantage of the least advantaged. Determining whether the Difference principle has been realized will be difficult. For, counterfactuals about how other possible distributions would ultimately impact the least well-off will standardly be quite opaque. Still, even if a conception of justice cannot facilitate an agent-guiding function in all instances, it could still be valuable if it can do so in some instances.

3.9 Political feasibility

For the remainder of this chapter I highlight the importance of the agent-guiding function of justice that I have disclosed by tracing out its implications for the work of prominent proponents of nonideal methodology. In this section I pave the way for this endeavor by demonstrating that my analysis applies mutatis mutandis to considerations of political feasibility.

As Pablo Gilabert and Holly Lawford-Smith note, ‘[r]oughly, a state of affairs is feasible if it is one we could actually bring about.’ Different conceptions of political feasibility

---

Pablo Gilabert and Holly Lawford-Smith, “Political Feasibility: A Conceptual Exploration,” Political
are found in the literature because philosophers disagree about what factors are relevant for determining whether we could actually bring about something. At one extreme, it could be argued that the only factors that are relevant are “hard” constraints: something is political feasible if and only if it is logically possible and compatible with the fixed biological limitations of humans. Alternatively, and more plausibly, it could be argued that what Gilabert and Lawford-Smith term “soft” constraints are also relevant factors. Such constraints include the economic, institutional and cultural (including religious) impediments to political progress. While soft constraints, unlike hard constraints, do not make it impossible for something to be brought about, they can make it very unlikely.

I will not propose a theory of political feasibility. Rather, I will argue from the following relatively non-controversial premise: our all-things-considered judgments, in nonideal conditions, ought to be at least partially responsive to the probability that candidate policies and laws will succeed, given the existence of soft-constraints. Suppose, for example, that the demands of justice dictate that women should have access to equal employment opportunities. In a particular set of nonideal conditions, however, the relatively widespread internalization of a set of cultural attitudes that view women as “naturally” domestic make the effective realization of this demand—at least in the short term—very difficult or impossible. It may be sensible for policy-makers to implement an intermediary policy to overcome

---

34 As Juha Räikkä notes, there are not just different conceptions of political feasibility in the philosophical literature. Other academic disciplines (i.e., engineering, economics, and political science) use the term differently as well. (See Juha Räikkä, “The Feasibility Condition in Political Theory,” The Journal of Political Philosophy 6 (1998): 27-40, pp. 28-29.)


37 Of course, it would be coherent to argue that our all-things-considered decisions should just reflect the demands of justice simpliciter, subject to hard feasibility constraints. This “pure” view has not been popular in the recent literature because it would require us to accept the conclusion that we should simply implement whatever policies and laws reflect justice simpliciter even if these have very negative consequences that undermine the short and long-term realization of justice. If this position is adopted, then the upshot of my argument against EA will not apply mutatis mutandis to considerations of political feasibility. However, if political feasibility is explicated in this very minimal sense, it is such a weak constraint that it is deprived of much significant philosophical interest.
some of these cultural attitudes and make the long-term goal of equal employment opportunities achievable.\(^{38}\)

Above I argued that in order for agents to adopt the right responses to institutional realizations of justice that reflect their epistemic limitations, it is necessary for them to have a conception of justice that is graspable independent of those limitations. Analogously, I argue that in order for agents to adopt the right responses to institutional realizations of justice that reflect soft-feasibility constraints, it is necessary for them to have a conception of justice that is independent of those soft-feasibility constraints. This is because agents should view arrangements designed to account for soft-feasibility constraints not as fulfilling the demands of justice, but rather as deviations from the demands of justice necessary for the effective long-term realization of justice.

Agents can view such feasibility-informed arrangements as deviations from the demands of justice if and only if agents have a conception of justice that is independent of such feasibility considerations. To be sure, not all actual agents will in fact view these arrangements in this way. For instance, the individuals who have internalized sexist values because of their exposure to unjust cultural institutions will not appropriately view arrangements designed to counter-balance those sexist values. However, an agent who is appropriately situated and responsive to the demands of justice will view these arrangements

---

\(^{38}\) An important question is what means are and are not acceptable to overcome soft-feasibility constraints in order to realize the demands of justice. This question is particularly difficult for deontologists, such as Rawlsians. For instance, as Samuel Freeman argues, a necessary condition for the full realization of ideal, Rawlsian principles of justice is that the principles must be public. This is because publicity is a precondition of moral and rational autonomy: publicity is necessary for agents to plan their actions in full light of the demands of justice. (Samuel Freeman, “The Burdens of Public Justification: Constructivism, Contractualism, and Publicity,” *Politics, Philosophy & Economics* (2007): 5-43, pp. 5-7.) Rawlsians must consider the extent to which this publicity condition should apply under nonideal conditions, and they may well be pulled in different directions. On one hand, the autonomy facilitated by the publicity condition is central to Rawlsian philosophy. Therefore, there is reason to think that the publicity condition should also constrain all policies in nonideal conditions. On the other hand, suppose that the only way to overcome certain types of soft-constraint that impede the realization of justice (e.g., the internalization of sexist values) is to introduce policies that do not meet this publicity constraint (e.g., to introduce policies aimed at eroding sexist values, in a way that does not disclose the reason for this erosion). Then perhaps, in order to be fair to the victims of the injustice that springs from these soft-constraints, we should temporarily suspend the publicity condition (at least under certain parameters). This difficult issue, which I will not here attempt to resolve, deserves more thorough exploration.
in this way. Therefore, an adequate agent-guiding conception of justice should idealize away both the epistemic limitations and soft-feasibility constraints that apply in the context of policy and law selection in nonideal conditions.

Interestingly, the political feasibility literature is on the edge of discovering this conclusion. In a recent paper Nicholas Southwood takes up the puzzle of whether “ought” implies “feasible.” He argues that intuitively we are deeply torn: on one hand, it seems as though if something is completely infeasible it is not something that we ought to do; on the other hand, it seems as though normative claims can be true in spite of the fact that they make infeasible demands. In contrast to other philosophers, Southwood takes this ambivalence at face value and argues that our normative talk encompasses a plurality of oughts.\(^\text{39}\) He argues that when we make normative claims that are independent from considerations of political feasibility, we are using (what he terms) the hypological ought: ‘Claims involving the hypological ought are supposed to be capable of operating in the service of our hypological practices: our practices of directing criticism towards others and ourselves.’\(^\text{40}\)

However, the hypological ought is only one type of normative discourse facilitated by (what I have termed) the agent-guiding function of justice. It is not only the case that we should criticize people for not performing acts that it is infeasible for them to perform at a particular time—perhaps, in order to achieve the broader political aim of helping to effect a social transformation. Further, as my previous examples establish, the agent-guiding function of justice can also prompt both the practices of not blaming policy makers because these policy makers have made the appropriate concessions to feasibility, and of acknowledging the social standing of people when the appropriate concessions to feasibility have been made.

I have also, in contrast to Southwood, provided a justification for accepting a type of normative demand that is independent from soft-feasibility constraints that goes beyond merely taking our ambivalent intuitions at face-value. These normative demands allow agents to respond in the correct way to institutional arrangements that are appropriately attuned to such considerations of feasibility.


3.10 Implications for prominent proponents of nonideal methodology

The fact that an adequate agent-guiding conception of justice must idealize away both epistemic limitations and soft-feasibility constraints enables new criticism of the charge that ideal theory is unnecessary. Sen, the foremost exponent of this charge, argues that the theory of justice that is valuable in nonideal conditions is “comparative”: a ranking of feasible societal arrangements. Crucially, on this comparative approach, it is not necessary to antecedently determine an ideal theory of justice—a fact that distinguishes Sen’s position from that of Rawls. It is, essentially a pairwise exercise: all we need to know is whether one feasible societal arrangement is preferable to another.41

3.10.1 The need to supplement Sen’s comparative approach with an idealized conception of justice

Assuming that my central argument has succeeded and granting for the sake of argument that Sen’s comparative approach is defensible, Sen’s approach must be supplemented with a further conception of justice. This conception will idealize away the epistemic limitations and soft-feasibility constraints that apply to actual agents, so that the outputs of Sen’s comparative approach can be viewed appropriately.

Of course, it would be premature to conclude from this that Sen must, in turn, acknowledge that nonideal theory depends on ideal theory. In order to block this claim, however, Sen must develop an adequately idealized conception of justice (i.e., a conception that idealizes away epistemic limitations and soft-feasibility constraints so that it can perform the agent-guiding function) that is distinct from an ideal (i.e., perfect) conception of justice.

I do not have the space here to determine whether this project can be achieved. There is reason to suspect it may be hard to achieve. For, as I have argued, one function that an agent-guiding conception of justice performs is the acknowledgment of the social standing of people in scenarios in which policies designed to improve their welfare fail. I suspect

that in order to perform this function adequately, it will not be sufficient to have norms that specify what justice would have demanded absent certain epistemic and feasibility considerations in a particular context (e.g., if a significant proportion of the population had not internalized sexist norms we would have implemented an equal employment policy). Rather it may require appealing to more abstract egalitarian norms about citizens' equal standing. If these more abstract norms are required then this starts to look very much like an ideal theory of justice and a corresponding set of nonideal principles of justice.

3.10.2 Does this value of justice lie outside the domain of the political?

A philosopher who wants to adopt a position at least roughly like Sen’s comparative approach to justice, and who wants to downplay the significance of the agent-guiding function of justice that I have uncovered, may argue that the agent-guiding function of justice lies outside the domain of the political. For instance, the distinctions that Gerald Gaus draws in his recent book *The Tyranny of the Ideal* would not allow this function to be accommodated within a practically valuable political philosophy. Gaus associates a conception of ideal theory that is not practically valuable with the work of David Estlund. Gaus writes that Estlund defends a “hopeless” view of justice, according to which ideal justice ‘...may be the true theory, even if it has no practical value.” In contrast, Gaus views his own project as one that rescues justice from uselessness by requiring ideal theory to perform the function of practical orientation. He writes that ‘...the ideal serves as a criterion that assists in regulating, directing, or facilitating less-than-ideal judgments of comparative justice.” In order to perform this orientating function—and to avoid endorsing Estlund’s “hopeless” view of justice—Gaus argues that an ideal theory of justice must satisfy (roughly) the following necessary “orientation” condition: it must play a *sui generis* role in ranking

---


feasible policies within nonideal conditions and, thereby, perform the practical function of
determining which policies ought to be selected.44

In Gaus’s conceptual cartography, a stipulative definition of what a practically
valuable political philosophy must achieve does a great deal of work. It is true that the agent-
guiding function of justice directs the attitudes and actions of agents rather than the institu-
tional selection of policies. Yet it is not as though this agent-guiding function directs the
private conduct of citizens when it comes to issues such as the virtues they should strive to
exercise in friendship; rather, the agent-guiding function of justice concerns the attitudes and
actions that agents should adopt in response to institutional arrangements.

I will register the fact that I would find the classification of the agent-guiding func-
tion of justice as apolitical eccentric. However, nothing of substantive importance ultimately
hinges on how philosophical terms of art such as “a practically valuable political philosophy”
and “orientation” are construed. Indeed, I have self-consciously presented my argument in
terms of “instrumental value” and “an agent-guiding function” to avoid entering into a
merely verbal dispute with philosophers such as Gaus.

The substantive problem with Gaus’s conceptual cartography is that he presents us
with a false dichotomy: either an ideal theory of justice must satisfy his necessary orientation
condition or—if it does not—it is at most, as Estlund argues, a conceptual truth about the
world that is intrinsically valuable, in virtue of the fact that it is true.

Suppose, though, that an ideal theory of justice does not satisfy Gaus’s orientation
condition because what realizing this ideal of justice would require is too epistemically
opaque. Even then, this ideal theory (if supplemented by the account of nonideal principles
of justice) could perform the agent-guiding function of helping agents to adopt the right
response to institutional arrangements. This would be a significant instrumental value of
justice, albeit—if Gaus’s cartography is accepted—an ethical rather than a political value.

3.10.3 Final comments

Even if the demands of nonideal principles of justice are, in some nonideal condi-
tions, so epistemically opaque that they cannot perform an institution-guiding function, it

44 Ibid. 40. (See Gaus’s formulations of what he terms the “Social Realization Condition” and “Orienta-
tion Condition.”)
would be wrong to conclude that they are inert. For those nonideal principles of justice can allow agents to adopt the right attitudes and actions in response to institutional arrangements and their subsequent—frequently unanticipated—social effects.

3.11 Conclusion

In this chapter I have outlined and defended the regulative function that the nonideal principles of justice perform in actual nonideal conditions. I have also cautioned that the instrumental value of justice in nonideal conditions cannot be reduced to this function. They also perform an agent-guiding function. An appropriate appreciation of this function has implications for prominent proponents of nonideal methodology.
Chapter 4

ANIDEOLOGICAL CRITIQUE OF NONIDEAL METHODOLOGY

4.1 Introduction

A distinctive feature of the nonideal theory of justice that I have developed in the previous chapters is that it accords a central function to ideal theory: the nonideal principles of justice constitute the transitional path to the realization of ideal justice.

Many contemporary philosophers would reject this approach because they endorse a nonideal methodology, which dispenses with ideal theory. A major reason why a number of such philosophers endorse this methodology is that ideal theory has been contested on the ground that it is ideological: essentially, that it performs the distorting social role of reifying and enforcing unjust features of the status quo.¹

In this chapter I argue that this ideological critique of ideal theory does not, in fact, provide good reason to ditch ideal theory in favor of a nonideal methodology. I do not

directly challenge this critique. Instead, I show that a parallel—equally compelling—ideological critique of nonideal methodology can also be presented. I argue that the rejection of the orthodox ideal theory paradigm can be explained by the increasing infiltration of capitalist and managerial social attitudes into academia. These social attitudes have commodified our conception of justice and consequently induced suspicion of ideal theory, which is not construed as having direct practical value. Nonideal methodology performs the distorting social role of reifying and enforcing unjust features of the status quo: the hegemonies of capitalism and managerialism that induced suspicion of ideal theory.

Proponents of the ideological critique of ideal theory are, consequently, confused about the ultimate dialectic upshot of their critique even if it succeeds: assuming that the method of ideological critique is philosophically viable, ideal theory and nonideal methodology are in symmetrically bad positions. As Karl Marx, himself, writes in the manuscript of the *German Ideology*:

_Everyone_ believes his craft to be the true one. Illusions regarding the connection between their craft and reality are more likely to be cherished by them because of the very nature of the craft…²

If this is correct then it is not surprising that proponents of the ideological critique of ideal theory have been so unreflective about the ideological forces that may be propelling their ideological critique and rival nonideal methodology.

My argument plays an important role in the argument of my dissertation as a whole. A view that has increasingly come to grip contemporary political philosophy—primarily because of the influential work of Charles Mills and Raymond Geuss—is that if we want to take the threat of ideological distortion seriously we should dispense with ideal theory. However, a hitherto unrecognized ideological critique of nonideal methodology can also be presented. Consequently, this view is untenable. If the threat of ideological distortion is taken seriously there is no reason to dismiss a nonideal theory of justice—such as my own—in which ideal theory performs a central function. Indeed, such a dismissal depends on incompletely confronting the spectre of ideological distortion.

This chapter has the following structure: I begin, naturally enough, with some preliminary clarifications. After that, I present a reconstruction of Mills’s ideological critique of ideal theory, before presenting my own ideological critique of nonideal methodology. I then anticipate and respond to a couple of objections, in order to further clarify and defend my position. I close by teasing out some general implications of my critique for contemporary work on ideological critique, and ideal and nonideal theory.

4.2 Preliminary clarifications

I begin with some preliminary clarifications about ideal and nonideal theory, and the method of ideological critique.

4.2.1 Ideal and nonideal theory

As I noted in chapter 1 (§1.3.1), the contemporary literature of ideal and nonideal theory is messy and indeterminate. A major reason for this is that the different characteristic features of ideal and nonideal theory—which capture how various philosophers have actually used the terms—cannot be combined into a unified conception of ideal and nonideal theory. This is because they can either come apart from one another or cut against one another. Recall that I tried to overcome this indeterminacy by defining the subject matter of ideal and nonideal theory exclusively in terms of characteristic features (1) and (6): in terms of a concept of how a perfectly just institution should be arranged and in terms of a conception of what we should do in conditions that fail to fully realize perfect justice.

I cannot, however, adopt this approach here, for an ideological critique needs to apply to an actual practice: onto how people use concepts, rather than onto a conception of how they ought to have used such concepts. The demarcation between ideal and nonideal theory has often been blurred; accordingly, the ideological critiques of ideal theory and nonideal methodology cannot be completely sharp. Nonetheless, it is important to recognize that there has been a substantive difference in the general approach between theorizing that could sensibly be labeled as ideal as opposed to nonideal: ideal theory tends to be abstract in form and to make claims about perfect justice; nonideal theory tends to be more concrete in
form and to make claims about how to ameliorate specific instances of actual injustices. This substantive difference in general approach is determinate enough for my present purpose.

4.2.3 Two elaborations of the claim that ideal theory is valuable

As I also noted in chapter 1 (§1.5.1), A. John Simmons and David Estlund offer different elaborations of the claim that ideal theory is valuable. Simmons, following Rawls, argues that ideal theory has indirect practical value: in nonideal conditions we should transition to the full realization of ideal justice, and this transition should be orientated by ideal theory. Estlund argues that ideal theory has intrinsic value that is independent of whatever practical value it may offer, such as its ability to orientate actual policy decisions. According to Estlund, ideal theory is important and it is valuable to come to understand something that is, itself, important.

In chapters 2 and 3 I offer my own development of Simmons’s claim. In this chapter I will consider the ramifications of ideological critiques for both of these different elaborations.

4.2.3 “Ideology” and “ideological critique”

Terry Eagleton outlines no fewer than sixteen different ways in which the term “ideology” has been used. Some of these uses are straightforwardly inappropriate to use in the context of an ideological critique of normative theorizing. For instance, as Raymond Geuss notes, ideology can be used in a purely descriptive sense, to denote the “world-view” of a

---


group that is systematically structured around a core set of interconnected beliefs.\textsuperscript{6} Alternatively, as John Levi Martin perceptively observes, ‘…many of us use it to denote the beliefs, attitudes and opinions of those with whom we disagree.’\textsuperscript{7}

I maintain that in order for a type or instance of normative theorizing to be classified as “ideological,” in the appropriate sense, at least one of the following two conditions must be satisfied. First, the theorizing plays a distorting social role\textsuperscript{8} of supporting unjust social practices or hegemonies—standardly, but not necessarily, the practice or hegemonies that propelled the theorizing. Tommie Shelby offers the best elaboration of this condition. He notes that ideology involves epistemic distortion. This epistemic distortion explains how ideology performs a distorting social role. For instance:

[a]ntiblack ideology wrongly claims that blacks are by their very nature inferior and thereby incapable of being the equal of writes…classic racist ideology fallaciously concludes that whites, in light of their alleged inherent superiority, should have a higher social and political status than that of blacks.\textsuperscript{9}

---


\textsuperscript{8} I use the term “social role” self-consciously because the term “social function” has become too loaded. This is largely because in an influential paper, Jon Elster attacks what he terms “functionalist explanations.” He argues that they postulate a purpose without a purposive actor and, thereby, presuppose a commitment to some form of objective teleology. Such teleological commitments cannot be reconciled with what Elster terms methodological individualism: ‘…the doctrine that all social phenomena (their structure and their change) are in principle explicable in terms of individuals—their properties, goals, and beliefs.’ (Jon Elster, “Marxism, Functionalism, and Game Theory,” \textit{Theory and Society} 11 (1982): 453-482, p. 453.) Neither my reconstruction of Mills’s ideological critique nor my own critique presuppose a commitment to objective teleology: such ideological forces can (at least in principle) be explained by the attitudes and interactions of actual people rather than the purposes of any independent phenomena (e.g., history or social institutions). Therefore, neither ideological critique is incompatible with methodological individualism.

Second, the theorizing is unlikely to be true because it is the product of distorting social attitudes that support unjust social practices or hegemonies. The caveat of “unlikely” is necessary to avoid the genetic fallacy: the fact that a piece of theorizing has a dubious causal history should induce suspicion about this theorizing but it does not entail that it is false.

These two conditions naturally harmonize with one another into the following claim: a set of distorting social attitudes give rise to theorizing unlikely to be true because it is the product of those attitudes. In addition, this theorizing performs the menacing social role of further reinforcing the social attitudes that propelled the generation of this theorizing. However, theorizing could plausibly be classified as ideological, in the appropriate sense, if it satisfied just one of these conditions. For instance, to focus on what I think is the most plausible possibility, it could be claimed that a piece of theorizing was playing a distorting social role even if the content of this theorizing is true.

It is important to note that the ideological status of theorizing is not (standardly) epistemically transparent to agents who are victims of ideological distortion. As Paul Ricoeur writes ‘[Ideology] operates behind our backs. We think from its point of view rather than thinking about it.’ Consequently, acting under the spell of ideology is distinct from acting in bad faith or as part of a conscious conspiracy.


11 Some philosophers might object that the genetic fallacy is sufficient to establish that an ideological critique of this sort is worthless: the causal origin of a set of beliefs does not entail anything about their truth-value. Consequently, we must turn to the assessment of arguments for the beliefs. Therefore, an ideological critique is redundant because it “takes us no further;” absent the ideological critique we would also simply have to turn to the assessment of arguments both for and against this set of beliefs. However, I think that this objection is (probably) wrong. The assessment of arguments for beliefs will (standardly) be a tricky business. An ideological critique could, therefore, play a substantial role in challenging these beliefs even if it does not straightforwardly entail that they are false. This is because, for instance, the ideological critique could provide *prima facie* evidence that a set of beliefs is false given their causal origin that is stronger than any argument that could be produced for why these beliefs are true.

12 Geuss carefully draws attention to this possibility in the following passage: ‘Since “masking social contradictions” might include such things as diverting attention from them, a form of consciousness might successfully mask social contradictions without containing any false beliefs.’ (Geuss, *The Idea of Critical Theory*, 18.)

Given that agents can fall under the spell of ideological distortion, the method of ideological critique is designed to perform a didactic function: to unmask the etiological source of the distortion, and to make explicit how the distortion perpetuates unjust social practices and hegemonies.\(^\text{14}\)

Some ideological critiques are explicitly tethered to Marxism. For instance, something can be construed as performing an ideological role if it ‘…hinders or obstructs the maximal development of the forces of material production.’\(^\text{15}\) This is not surprising given that the method of ideological critique emerged out of the Marxist tradition and the related critical theory tradition inaugurated by Max Horkheimer.\(^\text{16}\) Still, it would be a mistake to regard the method of ideological critique as dependent on the acceptance of substantive features of Marxism, such as historical materialism. The central claim underlying all ideological critiques—theorizing springs from and perpetuates distorting social attitudes—clearly does not depend on accepting Marx’s particular way of developing this claim into his base-superstructure model of explanation.\(^\text{17}\) (If the method of ideological critique were dependent on substantives features of Marxism then it would, of course, offer a dialectically weak form of criticism against non-Marxist theories.)

I close this section by emphasizing that my central argument does not presuppose that an ideological critique is a philosophically viable type of critique. I am just defending the following conditional claim: if it is a viable type of critique it supplies no reason to reject ideal theory and adopt a nonideal methodology.

4.3 An Ideological Critique of Ideal Theory

In what follows, I focus on Charles Mills’s ideological critique of ideal theory. Of all such critiques, Mills’s is the most well-developed and discussed version in the literature.

\(^\text{15}\) Ibid. 18.
Furthermore, as is crucial given my present purpose, Mills explicitly presents his critique as a decisive reason to favor a nonideal methodology.

4.3.1 Sociological claims

Mills’s argument builds from a couple of sociological claims. First, he states middle-to-upper-class white men ‘...are hugely over-represented in the professional philosophical population.’ Second, Rawls’s official justification for according theoretical primacy to ideal theory is that it is a necessary prolegomenon to the pressing and ultimately more important topics of nonideal theory. However, it is conspicuous that over his long writing career Rawls never really got around to addressing these pressing topics; contemporary Rawlsians have followed suit, largely evading such topics. Such an evasion, Mills argues, is not confined to the work of Rawlsians: in the large libertarian literature that dates back to the work of Robert Nozick, central topics of nonideal theory such as reparations for African Americans have hardly ever been discussed. Essentially, a preoccupation with ideal theory has meant that the urgent topics of nonideal theory have been ‘...perennially postponed to the tomorrow that never comes.’

4.3.2 An etiological and ideological hypothesis

Mills then extrapolates an etiological and an ideological hypothesis from these sociological claims. Given that throughout history philosophers have typically occupied privileged social positions, they have been uncomfortable with confronting the pressing topics of nonideal theory from which they have been personally insulated. Consequently, they have preferred to dedicate their intellectual attentions to ideal theorizing because this has allowed them to evade confronting these uncomfortable questions. Mills’s claim is not merely that ideal theory has this dubious etiology. Ideal theory is ideological because it also has the causal downstream effect of reifying and enforcing the social attitudes that propelled

---

19 Ibid. 178-180.
its invention and elaboration. Specifically, ideal theorizing induces a certain type of epistemic distortion: it conceals or “crowds-out” the really pressing issues of nonideal theory. This epistemic distortion explains how ideal theory performs the distorting social role of what could be called “privilege protection:” it serves the non-representative interests of privileged people who are overrepresented in academic philosophy by facilitating an evasion of pressing but uncomfortable questions of nonideal theory. In doing so it reifies and enforces unjust features of the status quo—for instance, a legacy of white supremacy that avoids confronting uncomfortable topics of nonideal theory, particularly racial justice.

In summary, Mills argues that the innovation of ideal theory springs from the distorting social perspective of privilege; ideal theorizing in turn further enforces that privilege by performing the distorting social role of privilege protection. As I noted above, the ideological status of theorizing is not (standardly) epistemically transparent to agents who are victims of ideological distortion. Accordingly, it is important to emphasize that the reason that ideal theorizing is in fact alluring has not (at least standardly) been accessible to the philosophers who have been preoccupied with ideal theory. They genuinely believed that ideal theory was intellectually and (perhaps also) socially valuable. However, Mills’s etiological and ideological hypotheses are arguably plausible explanations as to why ideal theory was invited, and the social role that it performs.

4.3.3 This ideological critique will (probably) also apply to nonideal principles of justice

I tentatively suggest that if Mills’s ideological critique of ideal theory succeeds it would also apply mutatis mutandis to nonideal principles of justice. This is because such principles have features that are characteristic of ideal theory: they are abstract principles constructed using the method of idealization. Consequently, the claim that, for instance, nonideal principles would “crowd-out” the pressing problems of nonideal theory would be as plausible as the claim that ideal principles of justice perform such a role. Of course, even if this suggestion is false, Mills's ideological critique of ideal theory still has direct implications for my nonideal theory of justice. This is because the realization of ideal theory is the ultimate goal of this nonideal theory.

4.3.4 Distinguishing this reconstruction of Mill’s argument from other claims

It is important to distinguish this reconstruction of Mills’s argument from other claims that are associated with—although independent from—this argument. First, it must be distinguished from an *ad hominem* objection to Rawls. In a paper dedicated to exploring Rawls and race, Mills writes:

Rawls was for decades at the most prestigious academic institution in the country…and, post- [*A*] *Theory of Justice*, as the book’s fame spread, he was the best-known and most celebrated political philosopher in the country…So why then could he—in the three decades up to his death, enjoying the success of *Theory*—not find the time to write even *one essay* on racial justice?22

The fact that Rawls did not directly confront the legacy of racial injustice in his work may perhaps reflect poorly on his intellectual and moral character. It does not, however, directly reveal any essential problems with ideal theory.

Relatedly, it must also be distinguished from a substantive criticism of a specific feature of Rawls’s ideal theory of justice. For instance, the claim that Rawls’s theory is purely forward-looking, and, therefore, it is poorly equipped to provide an adequate account of reparations for historic racial injustice.23 Again, problems with the actual content of Rawls’s theory of justice do not directly reveal any essential problems with ideal theory. In summary, an ideological critique of ideal theory must go beyond a criticism of either the character or work of Rawls.

Of course, the situation is a little tricky. In the twentieth century Rawls was the most famous exponent of an ideal theory of justice and (it is still probably accurate to say) Rawls’s ideal theory of justice (or a descendent of his theory) is still dominant. Therefore, the failings of either Rawls or contemporary Rawlsians might indicate a deeper *in principle* problem with the ideal theory paradigm. I simply emphasize that in the context of an ideological critique of

---

23 For a good version of such an argument see: A. John Simmons, *Boundaries of Authority* (Oxford: Oxford University Press, 2016), chapter 3.
ideal theory such objections are pertinent only in so far as they reveal such a deeper problem with ideal theory itself.

Second, an ideological critique of ideal theory must be distinguished from an ideological critique of a reflective-equilibrium methodology: the attempt to justify norms by achieving an equilibrium or harmony between particular judgments, general principles, and background social science. This methodology could be challenged on broadly ideological grounds as follows: we should not try to justify norms (e.g., a conception of justice) by seeking—in particular—to achieve an equilibrium that starts from (and attempts if possible to preserve) our particular judgments. This is because these judgments have been shaped by our immersion in a deeply unjust social world. In other words, the method of reflective-equilibrium gets things backwards: norms should not be extracted from our unjust social world, which is contaminated by legacies of injustice; rather norms that are independent from our unjust social world should be used to judge our unjust social world.

In Rawls's work there is a very tight connection between his ideal theory of justice and a reflective-equilibrium methodology because his ideal theory is constructed using this methodology. However, there is no essential connection between ideal theory and this methodology because an ideal theory could be constructed using a different method. Therefore, the ideological critiques of ideal theory and a reflective-equilibrium methodology come apart.

To clarify, with the exception of criticisms of Rawls’s character—which are best left to his biographers—I do not want to dismiss or downplay the philosophical significance of


25 Of course, this is not how a reflective-equilibrium methodology is officially supposed to work. Especially in what Norman Daniels calls “wide-reflective-equilibrium” our initial judgments about particular cases are just supposed to be provisional fixed points that must be critically scrutinized from every possible angle, including in relation to all alternative moral conceptions that have been developed throughout the history of philosophy. (See Norman Daniels, “Wide Reflective Equilibrium and Theory Acceptance in Ethics,” *Journal of Philosophy* 76 (1979): 256-282.) However, in response, someone pushing an ideological critique of a reflective-equilibrium methodology could emphasize that what is “officially” supposed to be the case is irrelevant: in practice this method seems, amongst other things, destined to reify unjust features of our social world.


27 See, for instance: Cohen, *Rescuing Justice and Equality*. 
any of these things that I have distinguished from an ideological critique of ideal theory. Indeed, in particular, I think that an ideological critique of a reflective-equilibrium methodology should also be explored. Of course such a critique—if successful—would have implications for my approach, which follows Rawls by embedding ideal theory in a reflective equilibrium methodology. However, given the focus of this chapter I simply want to pin my reconstruction of Mills’s ideological critique down to a critique of ideal theory per se.

4.3.4 Final comments on this ideological critique of ideal theory

My reconstruction of Mill’s argument can be classified as an ideological critique in the sense I outlined above. At the very minimum, Mills maintains that ideal theory performs the distorting social role of privilege protection. It is less clear whether he also wants to maintain that ideal theory is unlikely to be true because it is the product of distorting social attitudes. Mills—or another philosopher who wants to push an ideological critique of ideal theory—could reasonably make this additional claim. Alternatively, they could remain neutral about the truth-value of candidate ideal theories, whilst emphasizing that even if a candidate ideal theory is true it would (at least standardly) be performing the distorting social role of privilege protection and, therefore, concealing more important nonideal considerations of justice.

Philosophers who want to defend ideal theory really ought to confront Mill’s ideological critique. In this chapter I will only do so indirectly by constructing a parallel—equally compelling—critique of nonideal methodology. However, in the course of constructing and defending this parallel critique I will uncover some respects in which Mills’s argument needs to be modified. To foreshow a little, the claim that political philosophers have been almost entirely preoccupied with ideal theory and have almost entirely neglected the socially urgent questions of nonideal theory is an exaggeration. It is an accurate description of Rawls’s corpus of work but not of the whole of post-Rawlsian political philosophy. Therefore, a central sociological claim on which Mills builds his critique is actually weaker than he claims; accordingly, as I shall argue, the ideological hypothesis that one can rightfully extrapolate from that claim is weaker. Mills suggests that ideal theory performs a distorting role that is universal and complete; his ideological hypothesis would be more plausibly recast as the
claim that ideal theory performs—in some instances and to some degree—a distorting social role.

4.4 Inadequate responses to the ideological critique of ideal theory

Before I construct a parallel ideological critique of nonideal methodology, I will motivate the need to take this approach by surveying some other inadequate responses to this critique.

4.4.1 Tommie Shelby’s Rawlsian response

Shelby is one of the few philosophers who has tried to *directly* respond to Mills’s challenge, specifically on behalf of Rawls. Shelby writes that Rawls:

…is not vulnerable to the charge of implicitly endorsing white supremacy or the racial status quo. In fact, Rawls’s theory explicitly rules out regarding the members of any racial group as anything less than full moral persons.28

As Shelby subsequently goes on to explain, Rawls is explicitly clear that all legitimate conceptions of justice must satisfy formal constraints, such as “universality.” Shelby argues that this constraint of universality rules out discriminating against sub-groups of people on the basis of arbitrary factors such as race.29

Shelby’s analysis succeeds in establishing that the content of Rawls’s principles of justice do not have any racist implications. However, this is not sufficient to overcome Mills’s ideological critique. Even if the content of Rawlsian principles of justice do not have any racist implications, this does not mean that such principles are not being employed in an ideological capacity within actual social practices.

---


29 Ibid. 1702
4.4.2 The ideological critique of ideal theory is self-undermining

It is sometimes objected that the method of ideological critique is self-undermining.\(^{30}\) The objection runs as follows: if the ideological critique of ideal theory succeeds, it will also apply *mutatis mutandis* to this ideological critique because the claim that “ideal theory is ideological” is a claim about the viability of ideal theory. Therefore, it is, itself, a type of ideal theorizing. Consequently, it is, itself, also ideological.

This objection fails because relevant distinctions exist between ideal theory and the ideological critique of ideal theory; these distinctions prevent the easy application of this ideological critique to itself. The ideological critique is an extrapolation supported by sociological claims. It is, therefore, distinct from at least three of the characteristic features of ideal theory that I identified in chapter 1 (§1.2.1) in salient respects: it is not a perfect conception of justice, it is not a top-down approach, and no idealization is involved.

It is true that Mills’s critique consists of a very general—and indeed bold—claim about the methodological direction that *all* future theorizing should take. Someone could reasonably object to such a critique for exactly the same reason that she objects to ideal theorizing: it is intellectually hubristic for a theorist to try to make bold, general universal claims (whether about what perfect justice requires or about what type of theorizing should or should not be pursued). However, even if ideal theory and Mills’s ideological critique could be challenged using the same reason, this is distinct from the claim that Mills’s argument is directly self-undermining.

In order to clarify why this quick self-undermining objection fails, it is helpful to contrast Mills’s version of the critique with some different versions that are vulnerable to the self-undermining objection. Most obviously, if someone were to argue that all normative claims are ideological and we *ought* never to make normative claims, this argument would be straightforwardly self-undermining. More subtly, if an ideological critique of ideal theory is grounded in Marxism—as I argued above that it need not be—this ideological critique of ideal theory might be exposed to this quick self-undermining objection. Suppose, for

\(^{30}\) See, for instance: Ludwig von Mises, *Theory & History: An Interpretation of Social and Economic Evolution* (Auburn: Mises Institute, 1987), 126. (For further discussion, see: Shelby, “Ideology, Racism, and Critical Social Theory,” 154.)
instance, that a (somewhat crude)\textsuperscript{31} Marxist were to argue as follows: all norms are epiphenomenal and, therefore, ideological. This conclusion can be vindicated by the materialist conception of history, which is grounded in premises that as Marx, himself, maintains in the \textit{German Ideology}, can be ‘…verified in a purely empirical way.’\textsuperscript{32}

The claim that the ideological critique of normativity can be verified purely empirically is designed to block off the self-undermining objection to this ideological critique: the critique is not self-undermining because the critique—in contrast to that which is being critiqued—is purely empirical.

However, to put it mildly, post-Kuhnian philosophy of science has not been kind to the claim that this rigid demarcation between the normative and empirical can be maintained: social science, in particular, is not independent from social norms.\textsuperscript{33} Therefore, if this ideological critique of ideal theory were to succeed—which calls into question all norms—it would also apply to this ideological critique, which must implicitly presuppose norms.

In contrast, if like Mills, someone just argues that a particular type of theorizing (e.g., ideal theory) is ideological for particular reasons (e.g., because of an extrapolation from sociological facts) such a critique is not self-undermining.

\textbf{4.5. A Successful ideological critique of nonideal methodology}

In order to present a successful ideological critique of nonideal methodology, it is necessary to present specific reasons as to why the adoption of a nonideal methodology may be performing the distorting social role of supporting unjust social practices or hegemonies. Furthermore, in order to establish the thesis that the positions of ideal theory and nonideal methodology are symmetrically bad, the parallel ideological critique of nonideal methodology must be as plausible as Mills’s ideological critique of ideal theory. Essentially, although the self-undermining objection fails, the presentation of specific, equally powerful, reasons as to

\textsuperscript{31} I add the qualification “somewhat crude” because most scholars argue that Marx was not, himself, committed to the thesis that all normative theorizing is ideological. (Following Martin, “What is Ideology?,” 10-11.)

\textsuperscript{32} Marx, \textit{Collected Works Vol. 5}, 31.

\textsuperscript{33} This is particularly well argued by Helen E. Longino in her \textit{Science as Social Knowledge: Values and Objectivity in Scientific Inquiry} (Princeton N.J.: Princeton University Press, 1990).
why the adoption of a nonideal methodology maybe be performing a distorting social role would deflate the intended dialectical ambition of Mills’s argument (i.e., to show that we should dispense with ideal theory).

My ideological critique has an identical form to my reconstruction of Mills’s critique: I extrapolate an etiological and an ideological hypothesis from a series of sociological claims.

4.5.1 Sociological claims

_Capitalism_

Capitalism has had a deep and diffuse impact on education and research in the humanities. Most obviously, it has resulted in the commercialization of the humanities. The central reason for this is that politicians—no doubt responsive to popular sentiment rooted in a capitalist worldview—have extensively attacked the humanities on the ground that they are not a valuable means of promoting the economy. Barack Obama, for instance, has praised nations of the Far East, such as Singapore, in the following way: “They are spending less time teachings things that don’t matter, and more time teaching things that do.” As Nussbaum notes, when taken in context, “things that matter” is equivalent to “things that prepare for a career.” Similarly, former North Carolina Governor Patrick McCrory states: “[i]f you want to take gender studies that’s fine. Go to a private school and take it. But I don’t want to subsidize that if that’s not going to get someone a job.”

This prevalent criticism has set the terms for what would constitute a politically expedient defense of the humanities. Challenging the underlying capitalist assumption that the value of education can be reduced to its economic value is likely to fall on mute ears. Various defenders of the humanities have instead attempted to meet this criticism on its own terms.

36 Quoted in Tyler Kingkade, “Pat McCrory Lashes out Against ‘Educational Elite’ and Liberal Arts College Courses,” _Huffington Post_ (2013), New York: http://www.huffingtonpost.com/2013/02/03/pat-mccrory-college_n_2600579.html
For instance, the American Association of Colleges and Universities claims that the skills acquired through a liberal (humanities-based) education have become ‘America’s most valuable economic asset.’

Capitalism has also had a number of more subtle but deeply pervasive impacts that are causally connected to this increasing commercialization. First, it has resulted in what I call the “instrumentalization of value” in general. Members of the Frankfurt School, who refined and popularized the method of ideological critique, have extensively developed this point. For instance, in The Theory of Communicative Action, Jürgen Habermas tries to explain how capitalist modernization produces social pathologies. Most importantly, he argues that capitalism tends to colonize the lifeworld with instrumental reason.

In the educational context of the humanities, this instrumentalization of value has had a number of significant effects. Even if the value of the humanities is not always construed in the narrowest of commercial terms, it is increasingly viewed in purely instrumental terms. Nussbaum, for instance, has directly challenged the underlying capitalist assumption that education should be viewed as mere vocational training; yet even she defends the humanities on the instrumental ground that they contribute to democracy by creating good citizens.

The instrumentalization of value has not merely affected the way education is viewed in a broader social and political context. It has also had far deeper ramifications within the academy itself, with regard to the type of research that is valued and promoted. For instance, it is clear that philosophers face ever increasing pressure to direct their intellectual attention to “real,” “concrete” problems that “actually matter” and away from abstract philosophical theorizing that “does not matter” and is of interest only to philosophers themselves. Perhaps the most obvious example lies in the (in)famous social impact measures that all UK depart-

---


39 Nussbaum, Not For Profit, 6-10.
ments are obligated to submit at regular intervals.\footnote{For a good overview of this policy, see: Catriona Manville, “Measuring impact: how Australia and the UK are tackling research assessment,” The Guardian (2014): https://www.theguardian.com/higher-education-network/2014/dec/07/research-excellence-framework-2014-measuring-impact-australia-uk-universities-assessment} This practical shift is (at least partially) traceable to the instrumentalism wedded to capitalism: intellectual theorizing has value in so far as it can promote something outside itself.

Let us turn to a second pervasive effect of capitalism: the valorization of research that can be marketed as “important.” This effect is well documented by Gaus, who opens his polemical attack on applied ethics with the following observation:

‘Applied ethics’ is the growth industry in philosophy. Many philosophy departments, starved for students, have found a new, and apparently lucrative, market in teaching applied ethics. And many philosophers have found welcome research support in applying ethics.\footnote{Gerald Gaus, “Should Philosophers ‘Apply’ Ethics,” Think (2005): 63-68, p. 63.}

Obviously, writing papers on applied ethics does not have a commercial market any more than writing papers on mereology does. The relevant point, however, is that the increasingly “practical” bent of philosophy—illustrated by the growth of sub-disciplines like applied ethics—has its genealogy in the pressure to concretize research: it is this type of research that can be marketed to non-philosophers as important; as worthy of funding new hires, research grants, etc.

Managerialism

The infiltration of capitalist attitudes into every facet of academic life can partially explain our evolving conception of academic value. However, I argue that this shifting conception of value is also due to the distinct pressures exerted by what William F. Enteman and T. Klikauer term “managerialism:” the dubious doctrine that management is a self-standing, all-purpose expertise that can be applied to any organized human activity. Managerialism prizes the managerial methods of measurement, accountability, and audit. It also
propels the interests of managers and, thereby, deprives both workers and owners of control. As Bruce G. Charlton notes, the interests of managerialism can clash with those of capitalism. This is because ‘managers exert control by means of auditing…Therefore, organizations must be auditable – at any price…even when [this growth in auditing]…is exerting significant damage on performance and efficiency.’

In academia some of the changes that have resulted from managerialism are obvious: between 1975 and 2005 the ranks of administrators and non-faculty professionals nearly tripled. In doing so they have grown four times as fast as student enrollment. Given my present purposes, the most important change can be introduced by considering the following passage by Talbot Brewer:

These managers can exercise power only if the activity they manage can be conceived as something whose attainment they are in a position to ascertain and measure firsthand, without relying on the testimony of the trained professionals whose activities they are trying to regulate. Hence, their rise tends to be accompanied by a change in institutional aims and ambitions…

Managers lack specialized knowledge of the activities that they manage—academic managers of philosophy are not immersed in the intellectual activity of philosophy. Accordingly, they rely on a growing toolkit of measurements in the form of quantitative teaching evaluations, comparative rankings of institutions and journals, etc. that attempt to reduce the value of the activity to objective metrics that can be understood independently of the activity.

I also contend that managerialism induces a culture of suspicion toward any theorizing or research that can only be appreciated by becoming immersed in the intellectual activity

---

that constitutes such theorizing. The precise nature of this culture must be clarified with care. Technical science, mathematics, engineering, and applied ethics fall outside the purview of this culture in virtue of the fact that they are seen as promoting economic, technological, and social ends. And—crucially—the value of such ends can be appreciated without understanding the technical features of such research. In contrast, the value of research that seems disconnected from such ends is regarded as dubious, for it cannot be appreciated without becoming immersed in the activity that constitutes such theorizing.46

4.5.2 An etiological and ideological hypothesis

To begin my ideological critique of nonideal methodology, I will follow Mills’s approach by extrapolating an etiological hypothesis from these sociological claims. A plausible explanation as to why philosophers who propound nonideal methodology have rejected ideal theory is that these philosophers were both affected by and responsive to underlying social attitudes rooted in capitalism and managerialism—attitudes that have shaped the purpose of academic institutions, and also contorted and distorted the conception of intellectual value within such institutions. More precisely: their intuitions were shaped by the instrumentalization of value, and they were attracted to research that could be marketed as important, and could be deciphered as valuable by managers who were not immersed in this research. Simultaneously, they regarded research that could not satisfy the evaluative parameters set by the underlying social attitudes of capitalism and managerialism—research such as ideal theory—with suspicion.

In particular, the existence of hostility to the suggestion that ideal theory has intrinsic value is completely unsurprising. Such a claim stands in deep and irreconcilable tension with the instrumentalization of value in general. Furthermore, research into intrinsically valuable

---

46 It may be objected that this is not a bad thing: the claim that the value of an activity can only be appreciated by someone who is engaged in it is suspicious because it has been used to defend the value of a lot of pseudo-intellectual work and occult arts. This worry can be assuaged by appreciating the following important distinction: my claim is not—in contrast to many pseudo-intellectual movements and occult arts—that only “the initiated” who are engaged in and accept the value of a particular type of research can appreciate its value. Rather, my claim is that only people who have understood technical features of certain forms of research are in an adequate position to fully assess its value.
ideals is not the type of research that can be marketed as important. Nor can its value be apprehended without becoming immersed in the activity of philosophy in general and ideal theory in particular.

Also unsurprising is hostility to the suggestion that ideal theory is valuable because it should orientate nonideal theory. This mode of orientation is indirect; its application to nonideal theory is not something that ideal theory wears on its sleeve. Indeed, it has taken me the extended arguments of chapters 2 and 3 to show how ideal theory should orientate nonideal theory. Consequently, at least in contrast to nonideal treatments of specific problems of justice, it is neither something that can be marketed as important, nor something whose value one can apprehend without becoming immersed in the activity of philosophy in general and ideal theory in particular.

A crucial piece of data that supports this etiological hypothesis is the period of time during which opposition to ideal theory has galvanized such widespread support—the very end of the twentieth century to our present day: that is, a period of time in which capitalist and managerial attitudes have infiltrated academia to a degree of unprecedented depth.

As on my analogous reconstruction of Mills’s critique, my claim is not merely that the increasing adoption of a nonideal methodology has a dubious etiology. Nonideal methodology is ideological because it also has the causal downstream effect of reifying and enforcing the hegemonies that propelled its invention and elaboration. Nonideal methodology leads to epistemic distortion: it conceals the value of ideal theorizing. This epistemic distortion explains how nonideal methodology performs the social role of enforcing the ascendant hegemonies of capitalism and managerialism: it obscures the value of ideal theorizing that falls outside—and indeed potentially threatens—the evaluative parameters set by these hegemonies. In doing so, nonideal methodology reifies and enforces unjust features of the status quo: a culture rooted in capitalism and managerialism that has increasingly induced a contorted and distorted conception of academic value.

Recall that if Mills’s etiological hypothesis succeeds he must make the further claim that ideal theory induces the epistemic distortion of “crowding out” the urgent topics of nonideal theory, for it is this claim concerning epistemic distortion that grounds his ideological hypothesis. This epistemic claim is quite speculative because there is no a priori reason to think that ideal theory will have this effect, even though it could have such an effect in practice. In contrast, if my etiological hypothesis succeeds then the epistemic claim that
grounds my ideological hypothesis need not be speculative at all: we know a priori that a nonideal methodology will obscure the value of ideal theorizing because such a methodology explicitly rejects ideal theorizing.

4.5.3 Final comments on this ideological critique of nonideal methodology

As on my reconstruction of Mills’s critique, the reason why a nonideal methodology may in fact have been alluring is not (at least standardly) accessible to the philosophers who support such a methodology. It is nonetheless a plausible explanation for the increasing turn to nonideal methodology, and a plausible account of the social role that this methodology performs.

To parallel the ambiguity in Mills’s critique, I offer two different versions of my ideological critique: the first adds that the central claim of a nonideal methodology—we should dispense with ideal theory—is unlikely to be true because it is the product of distorting social attitudes. The second remains neutral about the truth-value of this central claim. It simply emphasizes that this nonideal methodology is performing a distorting social role of reifying and enforcing the hegemonies of capitalism and managerialism.

4.6 Objections and replies

I will anticipate and respond to a couple of objections, in order to further clarify and defend my ideological critique of nonideal methodology.

4.6.1 Objection 1

The ideological critique that I have presented of nonideal methodology is somewhat speculative. A range of alternative explanations can be given as to why ideal theory has been challenged. There is also at least some recalcitrant data that it is difficult for my critique to explain.
Reply

Yes, I completely agree. The increasing popularity of nonideal methodology might merely be temporally concurrent with the increasing infiltration of capitalist and managerial attitudes into academia. Laura Valentini, for instance, argues that ‘[i]n recent years, political philosophers have started to interrogate the methodology they use…’\(^{47}\) Perhaps this recent methodological self-consciousness has simply meant that more and more high quality arguments against ideal theory have come to light.

Also, there is some data that at least appears to be recalcitrant. Some of the most sophisticated arguments against ideal theory, such as David Wiens’s use technical models (e.g., Lewisian possible worlds).\(^{48}\) This type of methodological theorizing against ideal theory is difficult for my ideological critique to explain because it is perhaps even harder to “market” than much work on ideal theory. I grant that such data appears to be recalcitrant. However, for reasons that I will elaborate below, there is reason to think that even this technical work may be shaped by intuitions sculpted by social attitudes that have their etiology in the hegemonies of capitalism and managerialism.\(^{49}\)

I begin by re-emphasizing the central purpose of this chapter: to show that my ideological critique of nonideal methodology is as compelling as Mills’s ideological critique of ideal theory. To that end, the salient point is that these concerns about my critique also apply to Mills’s ideological critique of ideal theory with equal force. Mills’s ideological critique is also somewhat speculative, and other non-ideological explanations can be given for why Rawls and his early prominent critics such as Nozick focused on ideal theory. Most obviously, my central argument in chapters 2 and 3 that an adequate development of nonideal theory presupposes ideal theory might simply be correct.

---


\(^{48}\) See Wiens, “Against Ideal Guidance.”

\(^{49}\) An alternative response, which I will not pursue here, is as follows: Wiens’s research is connected to a research program closely associated with Gaus that uses the technical tools of economics. Economics is perceived as “useful” in broader society. Therefore, one way of responding to this allegedly recalcitrant data is to argue that Wiens’s work is sufficiently tied to this program that it could be marketed as “useful.” Consequently, this is not really recalcitrant data.
There is also recalcitrant data that is difficult for Mill’s ideological critique to explain; as I intimated above, his inaccurate narrative of twentieth and twenty-first century political philosophy plasters over this data. Mills is correct that Rawls and Nozick focused almost exclusively on ideal theory. However, many other political philosophers have made important contributions to both ideal and nonideal theory. 50 For instance, Thomas Pogge, who endorses a cosmopolitan version of Rawls’s ideal theory of justice, has addressed the topic of global poverty. 51 Joseph Carens has elaborated a novel egalitarian ideal theory, 52 and developed a theory of immigration that is appropriate for our actual nonideal world. 53 Philosophers who support a libertarian ideal theory have also engaged in extensive discussion about central topics in nonideal theory, such as reparations for historic injustices. 54 Clearly, the fact that an array of political philosophers have worked on both ideal and nonideal theory is hard to reconcile with Mills’s central claim that ideal theory has performed the distorting social role of postponing the urgent questions of nonideal theory to the tomorrow that never comes.

In summary, given that my aim is to present an ideological critique of nonideal methodology as compelling as Mills’s critique of ideal theory, it is unproblematic that my critique is somewhat speculative and faces recalcitrant data, because the same is true of his critique. However, in light of the recalcitrant data—with which both Mills’s and my critique must contend—I suggest that both critiques should be presented in a weaker form than Mills presents his critique. Mills should claim that ideal theory may well result in ideological distortion in some instances and to some degree, rather than that it does so in all instances and to an absolute degree. My critique of nonideal methodology must be analogously retained; it should claim that the rejection of ideal theory may well result in ideological distortion in some instances and to some degree.

4.6.2 Objection 2

If the central presupposition of my argument is true—that capitalist and managerial attitudes are responsible for deep tectonic shifts in academic culture—why is it that ideal justice in particular has come under attack? If my ideological critique were correct all abstract philosophical theorizing (e.g., metaphysics, meta-ethics, logic, etc.) would be equally under attack. Some additional explanation is required for why the subject of matter of justice has become the focal point of an ideal v. nonideal theory methodological debate.

Reply

I begin by reiterating that there has been a “practical shift” across the board—applied ethics, in particular, is the great growth area of philosophy precisely because, I suggest, it can be marketed to non-philosophers as something that “matters.”

The fact that justice has become the locus of this debate can be explained by noting that the subject matter of justice admits of two very different—almost opposite—types of philosophical exploration. First, it can be explored in a very abstract and theoretical way in complete independence from its application to actual political practice. In his late work G. A Cohen became famous for championing this approach. As he explains: ‘I want to know what justice is whatever I or anyone else may think is the right form and amount of the contribution that justice should make to political and social practice.’ Second, justice can be explored in a robustly practical way: as offering an account of a politically feasible institutional response that should be adopted to overcome a particular type of injustice, in a particular place, and at a particular time (e.g., gender based discrimination in Iran in the twenty-first century). Given that the subject matter of justice is uniquely theoretically flexible in this way, it is not surprising that it has become the focal point of a methodological debate.

In contrast, sub-disciplines like meta-ethics and mathematical logic will, by their very nature, always appear more abstract and devoid of robustly practical significance. Evidence that capitalism and managerialism also impact these disciplines can, I believe, be found in the

fact that the indirect vocational value of these disciplines is increasingly emphasized: for instance, that they help to build transferable analytic skills that employers value, improve scores on the LSAT, etc.

The status of contemporary metaphysics supports this general line of argument. Much of contemporary metaphysics seems abstract and devoid of robustly practical significance. That said, what Sally Haslanger describes as “ameliorative metaphysics” is enjoying increasing popularity: the attempt not merely to describe social kinds but to prescribe conceptions of social kinds which best serve the political end of promoting justice. A plausible sociological explanation for the rise of this ameliorative movement in metaphysics is that metaphysicians are being responsive to the evaluative parameters set by capitalism and managerialism: they are showing that their discipline can be marketed as addressing issues of social urgency and importance.

In summary, the current status not only of theorizing about justice but also of academic philosophy in general lines up (at least roughly) with what we would expect if my ideological critique of nonideal methodology is correct: applied ethics is insurgent, components of political philosophy and metaphysics that are capable of taking on a more robustly practical orientation have increasingly shifted in this direction, and purely theoretical parts of philosophy are either in comparative decline or are more and more self-conscious about their indirect vocational value.

4.7 General Implications

I will close by teasing out some general implications of my critique.

4.7.1 How to confront ideological distortion

If my central argument has succeeded, we arrive at a position that is tricky to overcome: ideal theory might be ideological, but it is just as plausible that the ideological critique of ideal theory, and the nonideal methodology designed to supplant ideal theory are ideological.

One available response is to argue that both critiques point towards Nietzsche’s famous observation about philosophy in general at the beginning of *Beyond Good and Evil*:

I have gradually come to realize what every great philosophy so far has been: a confession of faith on the part of its author, and a type of involuntary and unself-conscious memoir…

In other words, to conclude that all philosophy is ideological because any philosophical claim to objectivity is intellectually spurious: philosophical theorizing cannot escape the prison of the underlying social and psychological forces from which it emerges. This is the position that the sociologist Pierre Bourdieu occupies. He argues that all preferences—including intellectual preferences for different philosophical positions—are the products of the “habitus” (i.e., the economic and social determinants that structure experience) of different social classes. Consequently, ‘…philosophy in general is a groundless enterprise based on the fiction of pure, contemplative inquiry.’

This radical response is not, I think, warranted—at least not on the basis of either of the ideological critiques discussed in this chapter. As I have shown, both critiques are somewhat speculative and meet with some recalcitrant data. They are, therefore, not robust enough to justify rejecting philosophy *tout court*.

At the other extreme, it could be argued that the ideological critiques of both ideal theory and nonideal methodology are too speculative and thwarted by recalcitrant data to pass muster. This is a possibility that my central argument can accommodate. I am, after all, just defending the claim that the ideological critiques of ideal theory and nonideal methodology put them both in an equally bad position. Such a position might not actually be bad at all.

---


However I think that, at least prior to much more extensive critical scrutiny, we should not conclude that this possibility is true. After all, Mills is not making the anodyne point that ideal theory might be ideological because any mode of theorizing might be ideological. Rather, he has supplied a concrete reason as to why we should be suspicious of ideal theory in particular: the privilege protection role that it allegedly performs. Similarly, in my parallel critique of nonideal methodology I have endeavored to supply concrete reasons as to why we should be suspicious of nonideal methodology: the role that it performs in enforcing and reifying the ascendant hegemonies of capitalism and managerialism.

More generally, I think that an ideological critique is doomed by its very nature to be at least somewhat speculative. Diagnosing ideological distortion is perilously difficult. Given that ideological distortion distorts our theorizing, there is no way that we can step outside our embedded perspective and assume a meta-perspective from which we can once and for all infallibly determine whether we are victims of a particular type of ideological distortion. All we can do is to try to identify concrete reasons as to why ideological distortion may well have occurred in a particular context. If we can supply such reasons, then they should not be blithely dismissed for being “merely” speculative.

If this is correct, then whilst neither critique justifies rejecting philosophy tout court, the implications of both critiques should be taken seriously. To that end, we require some meta-strategies for attempting to overcome, to whatever extent possible, ideological distortion in both our theorizing and meta-theorizing (e.g., our theorizing about whether ideal theory is viable). Developing such strategies is obviously a large and difficult task, one I can only briefly touch on here.

Some of the recent literature on ideological critique has tried to devise meta-strategies specifically to prevent an ideological critique from, itself, being contested on ideological grounds. Significantly, if my central argument has succeeded, then one of the most prominent of such strategies is untenable. This prominent meta-strategy is defended by Robin Celikates, who writes:

---

The critique of ideology can...be reformulated as a specific case of the practice of critique without presupposing a privileged epistemic position and a break with ordinary practices of justification.61

As Haslanger notes: “[o]n his view, ideological critique is itself a social practice continuous with our everyday efforts to achieve reflective endorsement of our ongoing practices.”62

Celikates’s proposal has been appealing to philosophers such as Haslanger:63 in order to ensure that in presenting ideological critiques we do not, ourselves, tacitly adopt a position that is susceptible to an ideological critique, the aims of our critiques should be comparatively modest. We should not present our ideological critiques as external to ordinary social practice; rather, we should recognize that such critiques are embedded within such practices, and they are immanent in the sense that they (at most) reveal internal contradictions within such practices.64

However, this appeal stems from the fact that philosophers have hitherto failed to pursue an ideological critique of nonideal methodology. Celikates’s “modest” conception of ideological critique rules out the viability of appealing to ideal theorizing, at least in the context of such a critique. As I explained in chapter 1 (§1.2.1), ideal theory has characteristic features such as being (1) a conception of perfect justice, (2) constructed using the method of idealization, and (3) very abstract in form. These features make ideal theory discontinuous with ordinary social practices in important respects: ideal theory is not merely an imminent critique of social practices from an embedded social perspective—it consists (standardly) in an abstract claim about how social practices should be structured that is independent from the content of any actual practices. As I have shown, an equally powerful ideological critique of nonideal methodology can be constructed as an ideological critique of ideal theory. Consequently, a meta-strategy that does not allow room for ideal theory is not a viable

---

63 Ibid.
64 Following Titus Stahl, “Criticizing Social Reality from Within: Haslanger on Race, Gender and Ideology,” Krisis 1 (2014): 5-12.
means of preventing an ideological critique from, itself, being contested on ideological terms. This is because the rejection of ideal theory can, itself, be contested on ideological terms.

In contrast, the following two meta-strategies for trying to avoid ideological distortion in our theorizing seem promising, for they do not presuppose that ideal theory is not viable. First, there seems to be the need (at least when theorizing about social and political philosophy) to have a diverse group of theorists: when combined with the procedure of intellectual deliberation in good faith, this diversity will (with hope) bring to light instances and types of theorizing that result from ideological distortion. For instance, distortions that result from occupying a particular type of social position, such as a particular race, sex, class, nationality, age, sexuality, religion, etc.\footnote{This is what Justice Powell refers to as the value of a “robust exchange of ideas” in the context of his diversity based defense of affirmative action (\textit{Regents of University of California v. Bakke}, 438 U.S. 313.)} Second, as Haslanger notes, in order to overcome ideological distortion people often require new experiences, which unveil features of reality that were previously obscure to them and which, thereby, expose hereto held ideological distortions.\footnote{Haslanger, “Racism, Ideology, and Social Movement,” 10.}

4.7.2 Reassessing the value of ideal theory

Although an equally plausible ideological critique of both ideal theory and nonideal methodology can be presented, the implications of the ideological critique of ideal theory have been extensively rehearsed. In contrast, the implications of my ideological critique of nonideal methodology are uncharted territory. At the very minimum, this ideological critique of nonideal methodology should make us reassess the potential value of ideal theory in a new light.

Most obviously, it should give at least some pause to philosophers who are antecedently suspicious of the central aim of my dissertation project: to show that ideal theory—and idealized principles of nonideal justice—are central to nonideal theory. A hitherto unrecognized ideological critique of nonideal methodology can be produced. Consequently, an argument that the best account of what to do in nonideal conditions will require ideal theory should be carefully considered.
It also has implications for the debate between Estlund and Gaus that I introduced in chapter 1 (§1.5.1) and discussed further in chapter 3 (§3.10.2). Estlund argues that ideal theory has intrinsic value that is independent of whatever practical value it may offer, such as its ability to orientate actual policy decisions. According to Estlund, ideal theory is important and it is valuable to come to understand something that is, itself, important. Contra Estlund, Gaus acknowledges that it is possible that ideal theory may have intrinsic value even if it is not practically valuable. However, he derogatorily labels this a “mere dreaming account” to highlight that this is a very minor acknowledgment to make; an acknowledgment, he believes, that is compatible with overcoming the orthodox ideal theory paradigm.

Remarkably these two philosophers disagree about very little. Indeed, Gaus even acknowledges that Estlund’s central thesis—that an ideal theory of justice may have intrinsic value even if it is not practically valuable—may be correct. Their disagreement turns on the value that should be attached to the possibility of such intrinsic value. Estlund clearly thinks that it is philosophically significant; in contrast, Gaus thinks that allowing for the possibility of such intrinsic value is so insignificant that accepting the possibility is compatible with overcoming the orthodox ideal theory paradigm.

Given that a compelling ideological critique of nonideal methodology can be constructed, we should scrutinize the ideological forces that may be shaping Gaus’s underlying intuitions. Why should we accept, for instance, Gaus’s insistence that the ultimate vindication of ideal theory would be provided if such theory could perform the instrumental function of playing a sui generis role in ranking feasible policies within nonideal conditions? On what grounds is Gaus so sure that, if ideal theory cannot perform this function (for whatever reasons), its value is so ethereal that it is philosophically irrelevant?

---


70 See Gaus’s formulations of what he terms the “Social Realization Condition” and “Orientation Condition” in *Tyranny of the Ideal*, 40.
Interestingly, Gaus briefly speculates that his opposition to ideal theory results from a historical lesson stemming from the cold war. He notes that he counts himself immensely fortunate that Marxism (i.e., a dominating tyrannical conception of what ideal justice requires) and its threat to the liberal, democratic Open Society that he prizes is no longer a practical worry. He expresses indignation, however, that most academic philosophers have not been appropriately responsive to this bitter historical lesson because within the academy ‘the allure of the ideal is as powerful as ever.’\(^71\)

However, this is not really fair to prominent ideal theorists (e.g., Cohen, Estlund, Rawls, and Simmons) who are certainly not suggesting that an ideal theory of justice should be coercively imposed on actual people against their will. The philosophical value of ideal theory \(qua\) theory should be distinguished from its abuse in actual political practice. Once this distinction is made, Gaus’s “utopophobic”\(^72\) judgments about the value of ideal theory \(qua\) theory may well have their etiology in underlying social attitudes connected to capitalism and managerialism; furthermore, if this is correct, his philosophical position reifies and enforces these hegemonies.

4.7.3 Pessimistic and optimistic implications

As philosophers, we like to think that our inquiries are guided by over-arching norms such as truth, rationality, and reasonableness. The possibility that our theorizing is merely a reflection of underlying social attitudes—and, furthermore, reifies power structures that we sincerely believe ourselves to be opposing—is disquieting. I believe that the value of recent ideological critiques of ideal theory is that they impel us to confront this disquieting possibility. However, as I have argued in this chapter, we cannot evade the spectre that our

---

\(^{71}\) Ibid. xix.

\(^{72}\) This term is taken from Estlund, who defends the value of utopian political philosophy. Some of his most important arguments reveal that philosophers often wield the “ought implies can” dictum in crude ways that obscure salient distinctions. Perhaps most significantly, he argues, we should not confuse standards that are impossible to meet with standards that will certainly not be met (Estlund, Utopophobia, 199-120). In this chapter I have articulated an important component of a defense of utopian philosophy, which Estlund is yet to elaborate. This component is actually more suited to the title of “Utopophobia” that Estlund has, himself, introduced than Estlund’s own work. This is because it is an account of the social attitudes that propel a fear of utopian philosophy.
theorizing may be propelled by ideological forces simply by permanently shifting the subject matter of our inquiries away from ideal theory to nonideal theory. This simple solution is understandably seductive. Yet accepting it depends on incompletely confronting the threat of ideological distortion. For just as ideological forces may have infected the orthodox ideal theory paradigm that many self-styled nonideal theorists are striving to overcome, so ideological forces may just as plausibly have infected the ideological critique of ideal theory, and the nonideal methodology designed to supplant ideal theory. Therefore, the method of ideological critique has more pervasive implications than philosophers have standardly appreciated because it should be used to contest both ideal theory and nonideal methodology.

This might seem like a bleak implication. However, my analysis also points towards a more optimistic conclusion: it illuminates the appropriate, somewhat restricted, place of an ideological critique. Proponents of ideological critiques, such as Mills, standardly present such critiques as devastating and decisive. This, however, exaggerates their power. Mills’s critique of ideal theory is somewhat speculative and faces recalcitrant data, just as my parallel critique of nonideal methodology is also somewhat speculative and faces recalcitrant data. Therefore, although we should make every effort to scrutinize the ideological forces that may be propelling our theorizing, it does not seem rational, on the basis of either of these critiques, to abandon tout court any mode of theorizing.

4.8 Conclusion

In this chapter I have argued that the familiar ideological critique of ideal theory does not in fact provide a dialectically effective argument to abandon ideal theory in favor of a nonideal methodology. The threat of ideological distortion should be taken seriously but it cannot be overcome by abandoning ideal theory.
Chapter 5

NONIDEAL JUSTICE, FAIRNESS, & AFFIRMATIVE ACTION

5.1 Introduction

In this chapter I use the contractualist theory of nonideal justice developed in the previous chapters to specify the conditions under which non-meritocratic affirmative action policies are just. In doing so, I demonstrate how the major moral objection to affirmative action policies *viz.* that they are unfair can be overcome.

In applying some of the theorizing from the previous chapters to one of the most exigent issues of the present day I also have a broader philosophical aim: to illuminate the value of the nonideal theory of justice that I have developed. This task is important because if the theorizing of the previous chapters is viewed in isolation from such an application it may strike the reader as distressingly abstract—even scholastic—and devoid of significance. This unfortunate, although understandable, impression can be overcome by showing the value—perhaps even indispensability—of my theorizing for an adequate theoretical account of a pressing contemporary issue.

The chapter has the following structure: I begin by setting up the central moral puzzle concerning fairness and justice that non-meritocratic affirmative action raises. I then examine Elizabeth Anderson’s unsuccessful response to the objection, in order to illuminate what a successful solution must achieve. After this, I argue that the nonideal principles of justice that I have defended can be used to specify the conditions under which non-meritocratic affirmative action is justice, and to obviate the unfairness objection. I close by teasing out some legal implications of my argument, and reflecting in general on the value of my conceptual innovation of nonideal principles of justice.
5.2 Affirmative action

I will focus on affirmative action in the United States: a country in which such policies have taken on an enduring political relevance and controversy. In the United States, the term “affirmative action” was first used in an Executive Order issued by President John F. Kenney in 1961; the order included the stipulation that “[government] contractor[s] will take affirmative action to ensure that applicants are employed, and that employees are treated during employment, without regard to their race, creed, color, or national origin.” What taking affirmative action of this sort exactly requires is unclear. It is, therefore, perhaps unsurprising that the term “affirmative action” has since been used to label a disparate set of policies. It is illuminating to present the following taxonomy of affirmative action policies:

1. **Formal Equality of Opportunity**: the elimination of legal barriers to employment and education based on morally arbitrary characteristics (e.g., race and sex), and the punishment of private discrimination on the basis of such arbitrary characteristics.

2. **Aggressive Formal Equality of Opportunity**: self-conscious impartiality achieved through sensitivity training and external monitoring by government agencies (e.g., the Equal Employment Opportunity Commission) to supplement (1).

3. **Compensatory Support**: programs (e.g., special training programs, financial backing, daycare centers, apprenticeships, tutoring) to support members of target underrepresented groups. These programs are designed to compensate for the disadvantages that such members face because of discrimination and, therefore, to allow them to compete more effectively in hiring and admissions procedures.

4. **Soft Quotas**: non-explicit quotas (e.g., adding “bonus points” to university applications) for members of target underrepresented groups.

5. **Hard Quotas**: explicit quotas for members of target underrepresented groups, perhaps in proportion to the group’s representation in the wider population (e.g., if a third of the population is black a third of appointments in the police force must also be black).²

---


² A taxonomy of roughly this sort was first presented by Thomas Nagel. (See: Thomas Nagel, “Equal Treatment and Compensatory Discrimination,” *Philosophy & Public Affairs* 2 (1973): 348-363, pp. 349-51.) The
I will classify (1)-(3) as “weak” types of affirmative action policy and (4)-(5) as “strong” types of affirmative action policy. Weak policies are measures designed to facilitate greater access to education and employment for members of target underrepresented groups. Strong affirmative action policies use membership in target groups as a criterion for preference in the selection process itself.3

5.3 The appeal of affirmative action policies

Many actual societies are gripped by legacies of injustice. The United States is gripped—in particular—by an enduring legacy of racial injustice. The civil rights movement attempted to confront and overcome this legacy. Early affirmative action policies tended to be of a weak rather than strong sort. However, as Anderson notes, the removal of legal barriers was not sufficient to significantly increase the representation of target groups.4 Therefore, in the face of this failure, civil rights advocates increasingly came to support strong affirmative action policies.

The forward-looking justification for affirmative action policies in general is that they can be an effective means of (at least partially) overcoming legacies of injustice and, thereby,


3 Some philosophers have presented detailed conceptual analyses of “affirmative action.” (See, for instance: James P. Sterba, *Affirmative Action For the Future* (Ithica: Cornell University Press, 2009), 32.) However, as Kwame Anthony Appiah notes, “[t]he history of philosophy suggest that adequate definitions, conceived of as providing conditions individually necessary and jointly sufficient are, in any case, rather hard to come by.” (Kwame Anthony Appiah, “‘Group Right’ and Racial Affirmative Action,” *Journal of Ethics* (2011): 267.) This is particularly the case with the concept “affirmative action,” given that it has been used to categorize such a disparate set of policies; consequently, any candidate definition (in the form of necessary and sufficient conditions) threatens to be either uninformative or under-inclusive. I suggest, therefore, that it is more illuminating to define affirmative action ostensively, as I have done above.

advancing the cause of justice. The justification for strong affirmative action in particular resides, I believe, in the recognition that effectively confronting such legacies can require drastic courses of action, which go beyond simply removing unjust features of positive law.

5.4 The unfairness objection

The most powerful deontological objection to affirmative action is the unfairness objection. I begin by outlining the unfairness objection, before exploring its scope in the present day United States.


6 The so-called “color-blind” objection is the other major deontological objection: drawing distinctions on the basis of race are inherently wrong—even if they are not actually unfair on anyone. The color-blind argument can be undermined by demonstrating that the distinction drawn on the basis of race with respect to affirmative action policies is different from the distinction drawn with respect to race in racist forms of discrimination. As Elizabeth Anderson explains this can be done using the following analysis: ‘...[f]or a racial concept to figure in stigmatizing [i.e., racist] representations, it must add the idea that racial groups...differ in normatively significant character traits and talents as of intelligence, honest, sobriety, criminality, laziness...The concepts of race that figure in social scientific theories of categorical inequality, and in normative defense of...affirmative action...[do] not assert any intrinsic connection between race and normatively significant character traits.’ (Anderson, The Imperative of Integration, 156-7.) The distinction, therefore, is that racist forms of discrimination draw an intrinsic connection between racial features and a certain normative status—in racial forms of discrimination an inferior status. In contrast, race-conscious affirmative action policies do not depend on such a connection: it is not that race is intrinsically connected to a certain normative status, rather considerations of justice merely as a contingent matter of fact require us to draw distinctions on the basis of race in order to overcome instances injustice.
5.4.1  An Outline of the unfairness objection

Many critics of affirmative action policies argue that such policies are—at least in some instances of their application—unfair.\(^7\) The following example illustrates this charge of unfairness: imagine that a white man called Simon applies to Stanford Law School and is rejected. Further, imagine that the following counterfactual is true: if a soft-quota affirmative action policy of “adding points” to favor the admission of applications from members of target-underrepresented groups were not used, Simon would have been admitted. A proponent of the unfairness objection can grant that certain groups are underrepresented in some educational institutions and professions. However, such a proponent can simultaneously urge that people like Simon are not directly responsible for perpetrating the historic conditions that led to such underrepresentation. It is not, therefore, fair for the admission procedure to employ such a quota: this imposes the unfair burden of non-admission on Simon, simply because he has the same skin color and sex as the people who were responsible for perpetrating the present unjust conditions.

5.4.2  The relevance and scope of the unfairness objection in the present day United States

If we see philosophy as performing the public function of clarifying and enhancing the contours of actual public debate, the unfairness objection is significant because it is the objection to affirmative action that the general population tends to find compelling;\(^8\) further,


\(^8\) A recent poll, for instance, found that 88% of Millennials oppose affirmative action because they think that it is unfair. (See Matt Vespa, “MTV Poll: 88% of Millennials Think Affirmative Action Is Unfair.” http://www.cnsnews.com/mrctv-blog/matt-vespa/mtv-poll-88-millennials-think-affirmative-action-unfair.)
it is an objection with visceral appeal that motivates political indignation. In 2014, for instance, a group called Students for Fair Admissions filled a lawsuit against Harvard University, which alleged that Harvard *unfairly* discriminated against white and Asian students in their admissions procedure.\(^9\)

The unfairness objection can (plausibly) be applied if and only if the preference attached to membership in a target group results in the suspension of purely meritocratic features of admission and employment procedures. This is because it is the suspension of such features which gives rise to the charge that individuals, who are not member of the groups that such policies benefit, are being treated unequally hence unfairly.\(^10\)

The extent to which affirmative action in the United States today is in fact non-meritocratic and can, consequently, be (plausibly) contested on the ground of unfairness is the subject of ongoing dispute. In Frederick R. Lynch’s revealingly titled *Invisible Victims: White Males and the Crisis of Affirmative Action*, he claims that the unfair burdens of affirmative

---

\(^9\) For a good overview of this lawsuit against Harvard and the underlying political agenda of Students for Fair Admission, see the following article by journalist Michele S. Moses: “After *Fisher*: Lawsuit against Harvard goes on,” *New Boston Post* (2016). In the 2003 Supreme Court case of *Fisher v. University of Texas*, Fisher—a white woman—challenged the University of Texas at Austin’s use of race in the admission procedure, on the grounds that she was denied admission because of her race and that this, therefore, violated her right to equal protection under the Fourteenth Amendment. (*Fisher v. University of Texas* 570 U.S. 11-345 (2003).) Here the official ground of her complaint was constitutional. However, I strongly suspect that the reason that her predicament gained support from some of the public was because of the charge of unfairness that motivated the constitutional complain: like the example of Simon I devised above, it (perhaps) does not seem fair for her to face discrimination in the admissions process because of her race.

\(^10\) This conception of non-meritocratic affirmative action permits membership in a group to be treated as a criterion of preference in a selection process, if that treatment can be justified in purely meritocratic terms. Kwame Anthony Appiah and Amy Gutmann argue that this can be done to a (perhaps surprisingly) great degree. For instance, in an educational context, they argue that there is a case to be made for “diversity-role-models.” In teaching, for instance, white children to interact with non-white people on terms of equality and mutual respect, membership in another race is a qualification. (Kwame Anthony Appiah and Amy Gutmann, *Color Conscious: The Political Morality of Race* (Princeton N.J. Princeton University Press, 1998), 132.) If membership in such a group can be justified in purely meritocratic terms, this cannot be objected to on the ground that it is “unfair” with any more plausibility than someone can object that it is unfair that they did not get into an elite educational institutional because they are not endowed with sufficient natural intelligence.
action policies fall heaviest on white men. Interestingly, more recent research has concluded that, at least in the context of education, affirmative action policies impose heaviest burdens on Asian Americans. For instance, Thomas J. Espenshade and Alexandria Walton Radford found—even after controlling for factors such as grades, athletic status, etc.—that white people were three times, Hispanics six times, and black people approximately fifteen times more likely than Asian Americans to be accepted at a university in the United States.

Other research has concluded that the extent to which non-meritocratic affirmative action actually occurs is overblown. For instance, regarding affirmative action employment practices, a number of scholars have argued that—putting aside highly publicized examples such as the hiring of female firefighters—neither the purpose nor effect of most affirmative action policies has been to override purely meritocratic criteria. In summary, although it is largely accepted that there is some non-meritocratic affirmative action, the degree to which it in fact occurs is the subject of ongoing investigation and dispute.

I think that one major reason there is such extensive empirical dispute is that the mere fact that an affirmative-action policy is strong is not sufficient to ground a (plausible) charge of unfairness. After all, people’s explicit or implicit biases against members of an underrepresented group might be so strong that the only way to cancel out (or reduce the effect of) such biases is to implement a quota system. Alternatively, assume that the above counterfactual—concerning Simon’s happy admission fate absent a certain affirmative action policy—is true. Even so, this counterfactual cannot be considered in isolation. Other counterfactuals may also be true. For instance, if Simon were not the beneficiary of the structural advantages enjoyed by members of his privileged group stretching back several generations then perhaps he would not have been admitted. Essentially, affirmative action policies that cancel out an unfair advantage that Simon enjoys qua white male in the selection

12 Espenshade and Radford, No Longer Separate, Not Yet Equal, 1-13, 82, 431-5.
process are not unfair to Simon. (Moreover, not to cancel out that advantage is unfair to everybody who does not belong to his privileged group.)\(^{14}\)

Equally, as N. Scott Arnold explains, even some weak policies can plausibly be contested on the ground of unfairness. This is because such policies give rise to “disparate impact lawsuits”: anytime a company with fifteen or more employees has job requirements or hiring practices that have a negative (i.e., disproportional) “disparate impact” effect on minorities, it is risking a discrimination lawsuit. Therefore, the law may just officially require companies to implement weak policies that block discrimination. However, when such policies are viewed in their broader legal context, the prudent thing for some businesses to do—in order to ward off the potential expense of disparate impact lawsuits—is to implement a *de facto* quota system in their hiring processes.\(^{15}\) Such a *de facto* quota system could require suspending purely meritocratic features of the hiring process and, thereby, give rise to the unfairness objection.

I will not attempt to resolve the difficult empirical question of the extent to which non-meritocratic affirmative in fact occurs; I am concerned with the normative question of the extent to which, if any, non-meritocratic affirmative action is permissible.

5.4.3 Why the unfairness objection leads to an interesting philosophical challenge

As I explained above, affirmative action policies of all sorts have an obvious appeal: they can be an effective means of overcoming injustice. However, a *prima facie* plausible unfairness objection can be raised against non-meritocratic affirmative action policies—policies that may be an indispensable means of (at least partially) overcoming particularly engrained modes of injustice. Assessing the viability of non-meritocratic affirmative action is therefore tricky because reasonable judgments about justice and fairness pull in opposite directions. Consequently, non-meritocratic affirmative action is a good test case for an


adequate theoretical conception of nonideal justice. An adequate conception must specify
the conditions, if any, under which non-meritocratic affirmative action is just; a successful
philosophical defense must explain how the unfairness objection can be overcome. This
explanation must be intellectually persuasive as opposed to merely dismissive.

5.5. An unsuccessful solution to the Unfairness Objection

Anderson writes that the unfairness objection:

…neglects the fact that as long as discrimination or its effects persist, there
will be innocent victims suffering unjust burdens. The only question is
whether these burdens should be borne exclusively by disadvantaged racial
groups [and by extension disadvantaged groups in general] or more widely
shared. There is no injustice in sharing the costs of widespread injustice.16

Kwame Anthony Appiah enriches this line of argument by claiming that the burdens
imposed by non-meritocratic affirmative action policies are analogous to the burdens
imposed by other clearly permissible policies. He writes:

If justice requires restitution to Japanese Americans for the wrongs they suf-
fered in internment in World War II, I cannot complain, when my taxes are
raised to pay this restitution, that I did not do the interring.17

This argument presupposes that there is nothing in principle problematic with the government
instituting policies to distribute the burdens of dealing with the effects of injustice more
evenly, and that the demands of justice may impose certain costs on private individuals—
even if such individuals are not directly responsible for causing such injustice. These presu-
positions are highly plausible. This argument does not, however, establish that the particular
burdens imposed by non-meritocratic affirmative action policies are in fact just. After all,

16 Anderson, The Imperative of Integration, 139-140.
steadfast critics of non-meritocratic affirmative action policies, such as Louis Pojman, do not object on principle to the fact that just policies can impose certain costs on private innocent individuals. Rather, they object to the fact that non-meritocratic affirmative action policies impose particularly heavy burdens on select innocent individuals in the context of employment and education. Consequently, they conclude that such burdens are not in fact just.

Even if one is disposed to reject this conclusion, the imposition of such burdens is different from and more controversial than, for instance, the United States paying out reparations to the victims of state injustice and the cost of these reparations being evenly distributed amongst innocent taxpayers. In order to make this point clear, imagine—retaining Appiah’s example—that in order to compensate the interred Japanese Americans, a heavy tax was exclusively levied on a group of randomly selected non-Japanese people who comprised 5% of the population. This would distribute, at least in one sense, the costs of injustice more evenly. But it is an arrangement that the 5% group could plausibly contest as unfair—not because it imposed some costs on them but because it imposed particular costs exclusively on them.

This attempt to overcome the unfairness objection fails, then, because it faces the problem of what I call under-theorization: it does not provide a sufficiently rich and clear theoretical account of the demands of justice. The claim that the demands of justice can in principle impose certain costs is not sufficient to establish that a set of actual costs is in fact just. More subtly, even if this attempt were to succeed it would not be sufficient in and of itself to defend non-meritocratic affirmative action. This is because some further explanation would be required for why the fact that the burdens imposed by non-meritocratic affirmative action policies are just is sufficient to overcome the objection that they are unfair.

5.6 Rawls and Affirmative Action

A richer theoretical account is required. Accordingly, attempting to draw on the resources of Rawls’s monumental theory of justice seems like a promising approach to pursue.

---

As Pojman writes: ‘even if some compensation to blacks were indicated, it would be unfair to make innocent white males bear the whole brunt of the payments.’ (Pojman, “The Case Against Affirmative Action,” 108.)
I will argue that the nonideal principles of justice that I have derived can overcome the problem of under-theorization.

I pave the way for this endeavor by recapping some of the limitations of Rawls’s theory of justice as it stands, and examining Robert Taylor’s Rawlsian account of affirmative action. A careful reflection on Rawls and Taylor’s theorizing will point the way towards a more fruitful approach to Rawlsian nonideal theory: to use Rawls’s powerful innovation of a contractualist framework to forge *sui generis* nonideal principles of justice. This is, of course, the approach that I have undertaken in this dissertation—an approach that has been entirely neglected by other philosophers.

5.6.1 Rawls’s Ideal and Nonideal Theory of Justice

Rawls’s large corpus of published work contains little explicit discussion of the group-based injustices that affirmative action policies are designed to ameliorate (e.g., racism and sexism). That said, interpreters of Rawls—Tommie Shelby in particular—have shown that Rawls’s ideal theory of justice has substantive implications for such injustices. For instance, Rawls’s Liberty principle condemns race based-slavery and apartheid: under either institutional arrangement, *not* everyone would have access to the most extensive scheme of basic liberties that is compatible with other people’s exercise of such basic liberties. Similarly, the Fair Equality of Opportunity (FEO) principle condemns any employment or educational procedure that discriminates against a group, for such discrimination would by its very nature violate the terms of fair equality of opportunity. Essentially, from the perspective of Rawlsian ideal theory, these group-based injustices are objectionable because they are deviations from ideal justice.

As I noted in chapter 2, Rawlsian nonideal theory is transitional. Consequently, the goals of non-meritocratic affirmative action policies are good, at least in so far as they are an effective means of transitioning towards ideal justice. The crucial question is whether they satisfy the “permissibility” condition and are, thereby, an acceptable means of transitioning towards the ideal. However, as I also noted in chapter 1 (§1.4.2), the limitation of Rawls’s permissibility condition is that it provides no real guidance as to which transitional paths are

---

in fact permissible. Rawls’s conception of nonideal theory—as it stands—cannot determine whether non-meritocratic affirmative action in fact facilitates a permissible transitional path, for like Anderson’s approach it faces the problem of under-theorization.

5.6.2 An exposition of Taylor’s Rawlsian account of affirmative action

Given that neither Rawls’s ideal nor nonideal theory offer a neat or even adequate theoretical treatment of affirmative action, philosophers who wish to offer a broadly Rawlsian account need to be creative. Robert Taylor offers the most sophisticated Rawlsian treatment of affirmative action in the present literature.20 Taylor’s intricate argument can be re-constructed in the following four steps:

Step 1: Rawls’s nonideal theory must be constrained by his ideal theory. Otherwise it would collapse into consequentialism.21

Step 2: A necessary condition for nonideal theory to be constrained by ideal theory is that nonideal theory must not violate the spirit of ideal theory.22 The spirit of Rawls’s ideal FEO principle is captured by its pure proceduralism.23

---

20 Edwin L. Goff presents one of the first accounts of Rawlsian affirmative action. He clearly supports affirmative action in certain weak forms; however, towards the end of his article he writes that it is clear that both federal guidelines and partial compliance theory [i.e., nonideal theory] rule out hiring less qualified women and minorities over better qualified white males. (Edwin L. Goff, “Affirmative Action, John Rawls, and a Partial Compliance Theory of Justice,” Cultural Hermeneutics 4 (1976): 43-59, pp. 56-7.) It is not clear to me what precise gloss Goff wants to put on the crucial concept of “less or better qualified,” but it seems clear that this claim must be read as rejecting the viability of something at least roughly co-extensive with my concept of non-meritocratic affirmative action. If this is the case then Goff unilaterally rules out—with no real argument—the possibility that non-meritocratic affirmative action could be just under at least some conditions; consequently, his argument does not have any bearing on the central topic that I am examining in this chapter.

21 Robert S. Taylor, “Rawlsian Affirmative Action,” 488. (Taylor is explicit that this claim is strongly influenced by Christine Korsgaard’s interpretation of Rawls’s nonideal theory. (See: Christine M. Korsgaard, Creating the Kingdom of Ends (Cambridge: Cambridge university press, 1996), 147-51.)

22 The claim that nonideal theory must be consistent with the spirit of ideal theory is taken from Tamar Schapiro’s work, which extends Korsgaard’s interpretation of Rawls’s nonideal theory. (See: Tamar Schapiro, “Kantian Rigorism and Mitigating Circumstances,” Ethics 117 (2006): 32-57.)
Step 3: We must endorse pure proceduralism because ‘…we simply cannot know what the counterfactual result of a ‘clean’ competition would look like unless we run one…’\textsuperscript{24} This is because, for instance, under genuinely fair equality of opportunity there could be disproportional group outcomes that are (at least in part) due to cultural reasons not directly related to hiring practices: for example, Jewish overrepresentation among academics.\textsuperscript{25} Given the epistemic opacity of counterfactuals about what the outcomes of genuinely fair equality of opportunity would look like, Taylor argues that we ought to err on the side of caution. ‘At least for nonconsequentialist liberals, sins of commission should be of much greater concern than sins of omission—especially when the sinner is the state.’\textsuperscript{26} Therefore, at least in standard cases, we cannot tinker with the result of a procedure in order to bring it in line with an outcome that we antecedently judge to be fair.

Step 4: This analysis is applied to affirmative action. Weak affirmative action policies (e.g., equality of opportunity laws) are acceptable and moreover standardly required in nonideal conditions because they institute background measures that aim to make procedures as fair as possible. However, strong affirmative action policies (e.g., hard quotas) are almost always unacceptable. This is because we lack the epistemic capacity to judge what quotas would reflect the outcomes of genuinely fair equality of opportunity, and we should err on the side of caution.

5.6.3 Arguments against Taylor’s account

If Taylor’s argument succeeds then it will almost always be unacceptable to implement strong affirmative action policies—whether such policies are in fact meritocratic or

\textsuperscript{23} Taylor, “Rawlsian Affirmative action,” 490-493.
\textsuperscript{24} Ibid. 494.
\textsuperscript{25} Ibid. 498.
\textsuperscript{26} Ibid. 501
non-meritocratic.\footnote{Taylor suggests that a rare instance in which such policies might be justified in the present day United States is with respect to female fire fighters. This is because they are proportionally so underrepresented that we can be relatively confident if such policies are introduced (at least under certain parameters) that we are not overshooting the mark. (Ibid. 501-2.)} This conclusion will disappoint Rawlsians who wish to offer a more extensive defense of affirmative action. Of course, such disappointment does not constitute an objection to Taylor. He may be correct that central Rawlsian commitments lead to this conclusion.

In fact, as I shall argue, it is possible to provide a much more extensive Rawlsian justification of affirmative action. I will argue that this alternative, more extensive justification can be achieved whilst also preserving the central motivation and claim of Taylor’s argument.\footnote{For a more direct response to Taylor, see: D. C. Matthew, “Rawlsian Affirmative Action: A Reply to Robert Taylor,” \textit{Critical Philosophy of Race} 3 (2015): 324-343; Meshelski, “Procedural Justice and Affirmative Action.”}

The explicit motivation for steps 1 and 2 is that Taylor wants to prevent Rawlsian nonideal theory from collapsing into consequentialism. Here, I can be relatively brief and draw on my earlier argument in chapter 2 (§2.7.1). Taylor inherits the wrong conception of the relationship between ideal and nonideal theory from Korsgaard: nonideal theory is not \textit{constrained} by features of ideal theory. Rather, nonideal theory is \textit{transitional}. A false dichotomy motivates Taylor’s position. It is not the case that Rawlsian nonideal theory is either constrained by features of ideal theory and is, thereby, deontological; or is not constrained by features of ideal theory and is, thereby, crudely consequentialist. There is an alternative—not crudely consequentialist—conception of the relationship that I defended in chapter 2: the transition to ideal justice is subject to constraints of fairness—not merely concerned with realizing the goal of ideal justice as directly as possible. However, it is subject to such constraints of fairness because it is guided by \textit{sui generis} nonideal principles of justice, rather than because it is constrained by features of ideal theory.\footnote{To clarify, as I noted in chapter 2, in some respects the nonideal principles of justice have similar content to the ideal principles of justice. This is not surprising because like the ideal principles of justice they model considerations of justice and reflect the underlying values and predicament of the original-position contractors. However, the fact that they are similar in content should not make us infer the conclusion that} Consequently, although I reject
Taylor’s claim that nonideal theory is constrained by features of ideal theory, I preserve the deep motivation for this claim—to avoid collapsing Rawlsian nonideal theory into a crude version of consequentialism.

Taylor connects step 3 to the previous steps of his argument. However, step 3 does not depend on the success of the previous steps. The essential claim of step 3—our access to the requisite counterfactuals is relatively opaque and we should err on the side of caution—could be adopted even if the general claim that nonideal theory should be constrained by features of ideal theory, as well as the particular claim that one of those constraints should be the pure procedural spirit of the FEO principle are rejected.

I will grant for the sake of argument that the essential claim of step 3 is correct. However, I contend that Taylor is wrong to think that an adequate development of Rawls’s nonideal theory rules out affirmative action policies under a wide range of conditions because of the essential claim of step 3. I will argue that Taylor makes the mistake, as many Rawlsian and non-Rawlsian philosophers do, of thinking that the exclusive purpose of affirmative action policies must be to rectify injustice in education and employment procedures. This is a mistake because although affirmative action policies apply to such procedures, this does not entail that the exclusive function of such policies must be to overcome injustice within such procedures. Indeed, as I will argue, I think that the most compelling justification for non-meritocratic affirmative action policies is that they increase nonideal theory is constrained by features of ideal theory. This is a distortion of the conceptual relationship between ideal and nonideal theory and gives rise to fundamental problems with Taylor’s account.

I do have some reservations about this claim. In order to support his position Taylor draws an analogy with the criminal law and writes: ‘better to let the guilty go free (leave injustice uncorrected) than to punish the innocent (exacerbate injustice).’ (Taylor, Rawlsian Affirmative Action, 501.) I agree with this judgment about the criminal law; however, I do not reach it because I endorse the abstract principle that sins of commission should be of much greater concern than sins of omission. Rather, I reach this conclusion because I judge that the particular burden of imposing punishment on an innocent person is far worse than letting a guilty person go free. In contrast, it is not clear to me that this same type of judgment applies to affirmative action. While it may be bad for the state to overshoot the mark and impose unfair burdens on people, in the form of unjustified reverse discrimination, it is not clear to me that such imposition is worse than the state not instituting policies to rectify the effects of discrimination. “Reverse discrimination by the state” and “discrimination by private institutions” seem to me to be comparable harms; in contrast “punishing an innocent person” and “letting a guilty person go free” do not. Taylor’s argument hinges on an intuition that I do not share: that injustice perpetrated by the state is far worse than injustice perpetrated by private individuals.
comparatively disadvantaged people’s ability to exercise their basic liberties. In turn, this makes the requisite epistemic judgments far easier.

5.7 The nonideal liberty principle and non-meritocratic affirmative action

I will use the nonideal liberty principle that I derived in chapter 2 to specify the conditions in which non-meritocratic affirmative action is just. For ease of reference the principle is re-printed below:

I: Measures should be instituted to realize Rawls’s ideal Liberty principle of justice: ‘Each person is to have an equal right to the most extensive total system of equal basic liberties compatible with a similar system of liberty for all.’ The measures that are both permissible and required in order to achieve this aim are specified by the following clauses.

II: In determining what measures should be implemented in a particular set of nonideal conditions at time t, it is not sufficient to consider what measures would be the most effective means of increasing the standing of comparatively disadvantaged people at t. Rather, measures should be introduced that are the most effective means of increasing comparatively disadvantaged people’s standing prior to the full realization of ideal justice, regardless of which generation they are born in.

III: If comparatively disadvantaged people are disadvantaged to different degrees, priority should be given to the least advantaged.

IV: Comparatively disadvantaged people’s standing should be increased in such a way that it imposes as few demands on comparatively advantaged people as possible.

V: If it is not possible to increase the standing of disadvantaged people without imposing costs on comparatively advantaged people, such costs should be distributed according to the following two principles: (1) if possible, costs should be imposed on comparatively advantaged people; (2) the relative significance of these costs must be determined by the priority ordering of the
ideal principles of justice (e.g., the basic liberties have priority over equality of opportunity).

5.7.1 Tracing out the implications of this nonideal liberty principle for non-meritocratic affirmative action policies

Non-meritocratic affirmative action can be defended if the following conditions are satisfied:

(a) People who are comparatively disadvantaged with respect to their ability to exercise their basic liberties are in that position (at least partly) because they possess the characteristic(s) that non-meritocratic affirmative action policies are designed to target (e.g., the race or sex to which they belong).

(b) Non-meritocratic affirmative action policies are a generally effective means of increasing comparatively disadvantaged people’s ability to exercise their basic liberties, in a way that does not artificially privilege the interests of disadvantaged people at a particular time.

(c) There is not another candidate policy that would be as effective a means of increasing comparatively disadvantaged people’s ability to exercise their basic liberties; however, would impose less significant costs on comparatively advantaged people, as specified by the priority ordering of Rawls’s ideal principles of justice.

In order to try to determine whether these conditions are satisfied in actual societies such as the present day United States, we are called upon to make difficult empirical judgments. Clearly, the task of making such judgments cannot be left to philosophers alone. Rather, it is necessary to adopt an interdisciplinary approach and to appeal to rigorous social science. I will not try to undertake this requisite interdisciplinary approach here because my central aim is to defend the conditional claim that non-meritocratic affirmative action is justified if these conditions are satisfied. I will, however, briefly comment on each of these conditions and make a *prima facie* case that they might well be satisfied in the present day United States. After all, lacking such a *prima facie* case, the philosophical justification of such policies I offer would have no relevance for public policy in the present day United States.
That condition (a) is met seems comparatively uncontroversial: there is overwhelming anecdotal and statistical evidence that the enduring legacies of sexism and racism hamper individuals’ ability to exercise their basic liberties.\(^\text{31}\)

Determining whether condition (b) is met requires, of all the conditions, the trickiest empirical judgments. It rests on the following presupposition: insuring that placing a sufficient threshold of members of comparatively disadvantaged groups in certain educational institutions and professions will be a causally effective means of promoting members of such groups’ ability to exercise their basic liberties. There are a number of different ways in which this could happen. First, as Anderson argues, such thresholds could create greater social integration and break down the patterns of \textit{de facto} segregation in which group-based prejudices ferment. For instance, affirmative action policies help ‘…raise the probability that whites will have contact with blacks in large institutions, and thereby learn to interact more competently and comfortably with them.’\(^\text{32}\) Second, as Ronald Dworkin argues in \textit{A Matter of Principle}—especially on a large intergenerational scale—affirmative-action policies have the transformative social effect of breaking down negative race-based stereotypes. In particular, they can reduce statistical discrimination against black people by ‘…reducing the usefulness of race as a basis for making inferences about an individual’s position in society and conduct correlated with that position.’\(^\text{33}\) Daniel Sabbagh has provided extensive empirical evidence to support this claim.\(^\text{34}\)

Of course, whether affirmative action policies are effective in the requisite sense may not always be clear-cut. Even if a candidate affirmative action policy satisfies (b) for the above reasons, there may also be some associated costs. First, there might be the short-term effect of inducing stigmatization. For instance, Pojman writes that:

\(^{31}\) For a detailed but non-technical overview of the relevant statistical data, see: Anderson, \textit{The Imperative of Integration}; Sterba, \textit{Affirmative Action For the Future}.

\(^{32}\) Anderson, \textit{The Imperative of Integration}, 151.

\(^{33}\) This is how Anderson elegantly summarizes Dworkin’s position. (Ibid. 150) (See Ronald Dworkin, \textit{A Matter of Principle} (Cambridge, Mass.: Harvard University Press, 1985).

Fight the feeling how I will, I cannot help wondering on seeing a Black or woman in a position or honor, “Is she in this position because she merits it or because of Affirmative Action? Where Affirmative action is the policy, the “figment of pigment” creates a stigma of undeservedness, whether or not it is deserved.\footnote{Pojman, “The Case Against Affirmative Action,” 99.}

Note, Pojman is not merely observing that non-meritocratic affirmative action can create a stigmatizing effect that is proportional to the degree to which purely meritocratic considerations are suspended. Rather, he is highlighting that the practice of affirmative action in general—an umbrella term for both meritocratic and non-meritocratic forms—can create an effect that is disproportional to the degree to which non-meritocratic affirmative action in fact occurs. A second associated cost is that there might be a long-term backlash against the members of groups that affirmative action targets because people who are not members of such groups may wrongly believe that such policies are unjust even if the above conditions are satisfied.

Even if these costs arise to some degree, affirmative action could still be justified if these costs are outweighed by other benefits.\footnote{I accept that these costs have and will likely continue to arise, at least to some degree. In assessing these costs we must, however, be careful to consider whether affirmative action is the cause of such attitudes, or whether, for instance, racist attitudes that would have existed anyway simply latch onto affirmative action policies, perhaps in part as a justification for such attitudes.} However, given that clause II of the nonideal principle of justice stipulates that the interests of disadvantaged people at one time cannot be privileged over those of disadvantaged people at some other time, this may be sufficient (at least under some conditions) to rule out some affirmative action policies. Suppose, for instance, that a candidate policy was expected to impose sufficiently heavy burdens on disadvantaged people at time \( t \) because of the short-term effects of stigmatization. Such a policy cannot be justified, even if the policy would be a causally effective means of increasing the standing of different disadvantaged people at \( t+n \) in the long-term. Assuming, of course, that the people at \( t+n \) are not as disadvantaged as the people at \( t \). Essentially, candidate
affirmative action policies might be ruled out because they impose unfair burdens on disadvantaged people, given how actual people will behave.

Finally, let us turn to condition (c). According to the nonideal principle of justice that I have derived, it is clearly preferable to suspend purely meritocratic criteria in the context of employment and educational admission than to institute an alternative policy that imposes on comparatively advantaged people the cost of compromising the exercise of their basic liberties. This is because clause V stipulates that the significance of the costs must be determined by the priority ordering of the ideal principles of justice, and the ideal Liberty principle is lexically prior to the FEO principle.\(^{37}\)

However, the nonideal principle of justice that I have derived clearly does commit me to acknowledge that it would be preferable to pump money into, for instance, trying to combat racism, rather than to suspend purely meritocratic criteria in the context of educational admission and employment. This is because the FEO principle (which includes the proviso that jobs should be open to people on the basis of merit) has priority over the distribution of economic goods, according to the difference principle.\(^{38}\) Still, I suggest that we should be suspicious of the claim that “throwing money at the problem” will be sufficient to overcome the enduring legacy of—in particular—racism. For when we consider how socially engrained explicit and implicit biases against a group can be, it may well also be necessary to insure that a sufficient number of a racial minority are placed in certain professions and educational institutions to effect the deep social transformation that is required.

5.7.2 Three “good-making” features of my account

*Obviates the Unfairness Objection*

The unfairness objection gains critical traction because when non-meritocratic affirmative action policies are viewed in isolation from their broader context, they appear unfair: after all, particular individuals in an educational or employment context are not treated equally in terms of purely meritocratic criteria. However, this is not sufficient to


ground the charge of unfairness. For under the conditions that I have specified, non-meritocratic affirmative-action policies are fair because they secure a fair distribution of the burdens involved in transitioning to a more just world. The explanation for why they are fair can be presented using the nonideal contractualist framework: it would be rational for parties—who do not know what social position they will occupy, to assent to a principle that condones non-meritocratic affirmative action under the specified conditions. Therefore, the policy is impartial—hence fair—in the appropriate sense.

The topic of affirmative action instructs us to consider the relationship between “fairness” and “justice.” Many folk use these two concepts interchangeably. However, some philosophical theories of justice render these concepts completely distinct: for instance, a theory in which justice is explicated as “upholding people’s natural rights.” On such a conception of justice showing non-meritocratic affirmative action to be just would not be sufficient to obviate the unfairness objection. Some additional account would have to be supplied in order to explain why considerations of justice have priority over considerations of fairness.

In contrast, a Rawlsian contractualist framework is uniquely suited to overcoming the unfairness objection because it presupposes such a tight connection between the concepts of “justice” and “fairness.” Rawls describes his view as “justice as fairness” because “…the principles of justice are the result of a fair agreement or bargain.” It would be an exaggeration to say that in a Rawlsian framework, fairness and justice are simply synonymous concepts. Even so, if the conditions under which non-meritocratic affirmative action are just obtained, it is difficult to see how a plausible charge of unfairness could be presented against such policies. After all, they are justified by a principle that result from a fair agreement.

*Not Narrowly Focused on Equality of Opportunity*

Rather than being focused on the long-term goal of achieving genuinely fair equality of opportunity, the defense that I have provided for affirmative action appeals to its broader function of promoting people’s ability to exercise their basic liberties. This has a number of philosophical advantages.

---

39 Ibid. 11 (emphasis added).
First, my argument rests on a robust Rawlsian ground, for Rawls accords lexical priority to the basic liberties. This priority justifies suspending purely meritocratic criteria in the context of educational admission and employment procedures, if such suspension is necessary to promote people’s ability to exercise their basic liberties, because the FEO principle is lexically subordinate to Liberty principle. Even if priority rules—admittedly one of the most controversial features of Rawls’s theory of justice—are rejected, the nonideal contractualist framework can still be used to defend non-meritocratic affirmative action so long as basic liberties are accorded significant value. On such an approach, non-meritocratic affirmative action will not simply be justified if it is an effective means of promoting people’s ability to exercise their basic liberties; rather it will be justified if it is an effective means of promoting people’s ability to exercise their basic liberties to a specific threshold that does not impose unreasonable short-term costs on equality of opportunity.

Second, it allows my account to side-step one of the major objections to actual non-meritocratic affirmative action policies. Some social scientists have argued that “poverty” as opposed to “race” or “sex” is a much better metric of comparative disadvantage in educational admission and employment. Therefore, if something like non-meritocratic affirmative action is defensible it should be on the basis of class and not—as is standard—on the basis of race and sex. Even if it is true that relative poverty is a more reliable metric for comparative disadvantage in education and employment contexts, non-meritocratic affirmative action can still be justified if “race” and “sex” are more reliable metrics for comparative disadvantage with respect to an ability to exercise one’s basic liberties.

Third, my argument reduces a certain epistemic concern. To return to an earlier thread of argument, assume that Taylor is correct both that counterfactuals about the outcomes of genuinely fair admissions procedures are relatively opaque, and that we should err on the side of caution in tinkering with such procedures. On my account, these claims are

---


not, as Taylor urges, sufficient to rule out affirmative action policies that take the form of soft and hard quotas under a wide range of conditions. For the epistemic judgments required on my account are significantly easier to make in two crucial respects. First, innocent group preferences may make it hard to predict the output of genuinely fair admission and hiring procedures. However, in contrast, it is not plausible to think that innocent group preferences could account for the fact that members of some groups are able to exercise their basic liberties to a lesser degree. Essentially, the demands imposed by the Liberties principle are considerably less opaque than the demands imposed by the FEO principle. Second, basic liberties are accorded greater value than equality of opportunity; therefore, even if in general we should err on the side of caution, it is not clear that we should err on the side of caution if the potential harm applies with respect to equality of opportunity, whereas the potential benefit applies with respect to basic liberties.

*Can explain eccentric features of actual affirmative action policies*

As Tom L. Beauchamp observes, "[p]aradoxically, the younger minority members and women who have suffered least from discrimination now stand to gain the most from affirmative action."\(^{42}\) Antonin Scalia expands on this line of argument by noting that the *costs* may also fall on the least deserving, when he writes that affirmative action will:

…prefer the son of a prosperous and well-educated black doctor or lawyer—solely because of his race—to the son of a recent refugee from Eastern Europe who is working as a manual laborer to get his family ahead…\(^{43}\)

These features of actual affirmative action policies are difficult for backward-looking justifications to explain. However, on my forward-looking approach it does not matter if some—even all—of the recipients of affirmative action policies are comparatively advantaged either with respect to their ability to effectively compete in admissions procedures or their ability to exercise their basic rights. If such policies have the effect of helping to promote the social


\(^{43}\) Scalia, “The Disease as Cure: ‘In Order to Get Beyond Racism, We Must First Take Account of Race.”
standing of other members of these groups then they can be justified under the specified parameters.

5.8 Legal implications and a final remark about methodology

I will round out this chapter—and indeed my dissertation as a whole—by tracing out some legal implications of my argument, and by offering a more general methodological reflection about the value of the conceptual innovation of nonideal principles of justice.

5.8.1 Legal Implications

If the conditions under which non-meritocratic affirmative action is just obtain in the present day United States then, subject to constraints imposed by constitutional law, my argument offers compelling reasons for thinking that we should expand the use of affirmative action beyond the precedent set by the 2003 Supreme Court decision of *Grutter v. Bollinger*. This precedent prohibited educational institutions to use quotas—whether such policies are in fact meritocratic or non-meritocratic.

Equally, if my argument has succeeded, there is reason to reign in and modify features of actual affirmative action policies in the present day United States. As I noted above, there is empirical evidence that affirmative action, at least in the context of education, imposes heaviest burdens on Asian Americans. Suppose that this evidence is correct. If the purpose of non-meritocratic affirmative action policies is to promote comparatively disadvantaged people’s ability to exercise their basic liberties, there is no reason to think that the burdens of such policies should fall particularly heavily on Asian Americans just because they are comparatively well represented in higher education.

---

5.8.2 A closing reflection on the value of idealizations, particularly nonideal principles of justice, for nonideal theory

As I am sure is now apparent to someone who have read the whole of my dissertation, there is an increasingly voluminous literature dedicated to attacking the value of ideal theory in general and Rawls’s theory of justice in particular, given that he is identified as the paradigm exponent of ideal theory. As I acknowledged in chapter 3 (§3.2), some of these arguments—for instance the claim that abstract principles of justice determined in a Rawlsian contractualist framework cannot offer adequate normative guidance in the actual world—also apply mutatis mutandis to the nonideal principles of justice that I have defended. This is because nonideal principles of justice—like ideal principles—are idealizations that are determined within a Rawlsian contractualist framework.

Many of these attacks on ideal theory appear in the form of very general (at times polemical) claims about what ideal theory—given its form and method of derivation—cannot in principle achieve. They are also often accompanied, in turn, by an alternative grand vision of what political philosophy should aspire to achieve in order to be practically valuable.

It is worth pausing to note that there is a slight irony to this state of affairs: after all, I take it that the deep philosophical motivation for moving away from the Rawlsian paradigm of ideal theory is that such a paradigm is judged to be too narrowly focused on the project of making general abstract claims about justice and, consequently, is insufficiently attentive to concrete features of reality. And yet, the case for the rejection of ideal theory has itself standardly been made in incredibly general abstract terms.

Throughout this dissertation, I have argued that all of the radical skeptical arguments that I have surveyed against ideal theory fail. For instance, in chapter 3 (§3.2.2), I argued that the attempt to establish a priori that ideal theory is in principle incapable of offering adequate normative guidance in the actual world is unsuccessful. In chapter 4, I argue that whilst ideological distortion should be taken seriously, the familiar ideological critique of ideal theory fails to supply a dialectically effective reason to abandon ideal theory in favor of a nonideal methodology because an equally compelling ideological critique of nonideal methodology can also be presented.
Assume, for the moment, that all the radical skeptical arguments against ideal theory can be successfully repelled. I suggest that the methodological question of what types of theorizing (e.g., ideal theory) and concepts (e.g., nonideal principles of justice) are instrumentally valuable should be settled by using the following test: consider which types of theorizing and concepts are required to craft the most compelling philosophical accounts of first-order problems (e.g., When, if at all, is non-meritocratic affirmative action just?). Essentially, the second-order methodological question should be determined by what adequate treatment of the first-order problems requires.

In the context of racial injustice in particular, a number of philosophers, including Anderson, have rejected the ideal-theory paradigm in favor of a nonideal methodology. Anderson takes this approach for two reasons. First, following Amartya Sen, she argues that it is not necessary to work out an ideal theory of justice in order to offer an adequate nonideal treatment of racial injustice. Second, she argues that political philosophers—instead of wasting their time trying to antecedently determine very abstract ideal principles of justice—

---

45 I am, here, leaving open the possibility that ideal theory could have intrinsic value. This is the position that David Estlund defends. (See: David Estlund, “What Good Is It? Unrealistic Political Theory and the Value of Intellectual Work,” *Analyse & Kritik* 33 (2011): 395-416.)


47 Anderson, *The Imperative of Integration*, 3. (See also: Amartya Sen, “What Do We Want from a Theory of Justice?,” *Journal of Philosophy* 103 (2006): 215-238, pp. 218-222.) It is not at all clear to me, despite the way Anderson officially describes her project, that the argument she actually presents departs from the parameters of a broadly Rawlsian conception of ideal and nonideal theory. For instance, in a passage that offers a good overview of the book’s argument as a whole, Anderson writes: ‘The integrative model [i.e., the model for responding to race-based injustice that she defends] represents racial preferences [e.g., affirmative-action policies] as a means to racially integrate the main institutions of civil society. Segregation is the linchpin of unjust systematic race-based disadvantage…Integration helps dismantle these underlying causes of race-based injustice…Integration is also needed to advance a democratic culture, by providing opportunities for citizens from all walks of life to communicate on matters of public interest…This is a forward-looking rationale: it views the integration of mainstream institutions as essential to advancing justice and democracy.’ (Anderson, *The Imperative of Integration*, 136.) Whilst her approach is clearly more empirically informed that many broadly Rawlsian treatments of affirmative action, her conception of nonideal theory, like that of Rawls, is still transitional and presupposes an ideal theory of justice—the ideal of an integrated democratic culture.
should rather take an empirically informed “bottom-up” approach and attempt to respond to the concrete problems that they face at a particular time and place (e.g., racial injustice in the present day United States).  

As I have argued, Anderson is unable to overcome the unfairness objection because she faces the problem of under-theorization. This problem of under-theorization does not merely apply to Anderson’s position. Philosophers and political advocates of affirmative action have long recognized that the theoretical basis for defenses of affirmative action policies is insufficiently developed. For instance, the philosopher Ira Katznelson, echoing the sentiments of Jack Greenberg, the former director-counsel of the NAACP Legal Defense and Educational fund, writes: ‘Defenders of affirmative action typically argue with a body of principled reasoning appreciably less developed than that of the opposition.’

Assuming that I am right that the second-order methodological question should be determined by what is required to address first-order problems, then it is incumbent on philosophers like Anderson to show how a nonideal “bottom-up” methodology can provide an adequate response to the unfairness objection. I am not optimistic that this project can be successfully pulled off: given that the debate about non-meritocratic affirmative action hinges around such deep questions of fairness and justice, I think that something very like my conceptual innovation of nonideal principles of justice will be indispensable.

5.9 Conclusion

In this chapter I have argued that the nonideal principles of justice I have derived can be used to specify the conditions under which non-meritocratic affirmative action is just and to ameliorate the objection that such policies are unfair under such conditions. In doing so I have tried to vindicate the value of my work in this dissertation as a whole by showing the value—perhaps indispensability—of nonideal principles of justice for nonideal theory.


Appendix I

CONTRACTUALISM & INTERGENERATIONAL JUSTICE

I.1 Introduction

In this appendix I defend the “representation of all generations” model that I use in chapter 2. I begin by reflecting on the value of constructing a contractualist model, in order to illuminate what a viable model of intergenerational justice must achieve. I then examine and reject some prominent contractualist models of intergenerational justice. After that I defend the representation of all generations model, in part by explaining why it is an improvement of Rawls’s “general assembly” model. I close by anticipating and responding to a couple of objections.

I.2 The value of constructing a contractualist model

In order to build up a clearer sense of what is at stake, it is important to begin by considering the value of constructing a contractualist model. Such a model should function as an explanatory heuristic: it should explain the demands of justice by illuminating and clarifying them. In order to do this it must avoid certain things. Most obviously it must avoid being trivial. If the model is set up in such a way that the output of the contractualist procedure in the model immediately—and obviously—follows, then the model fails to perform any heuristic function. In order to avoid this pitfall, the model must leave genuine room for the contracting parties to deliberate and consider the costs and benefits of candidate outputs. Relatedly, the model cannot be so ad hoc that it is deprived of any real explanatory value.
I.3 The present time of entry interpretation

In *A Theory of Justice* Rawls (briefly) considers questions of ideal, intergenerational justice. He defends a just savings principle: the institution of putting aside capital for the benefit of future generations.\(^1\) Rawls stipulates that the contracting parties should know that they all belong to the *same* generation. He refers to this as the “present time of entry” interpretation.\(^2\)

The contracting parties know that they all belong to the same generation. They do not, however, know in which generation they will actually be born. It might, therefore, seem as if this ignorance is sufficient to facilitate the determination of intergenerationally just principles (e.g., a just rate of savings): in order to maximize their own interests the parties need to select a rate of savings that is not so high that it imposes excessive costs on them if they happen to be born in a relatively early generation, and not so low that they cannot accrue the benefits of early generations’ savings if they happen to be born in a later generation. Therefore, in order to rationally maximize their own interests they will select, for instance, a principle of savings that is in fact intergenerationally fair.

However, if the contracting parties want to rationally maximize their own interests they will not—at least as the contractual model stands—select a rate of savings principle that is intergenerationally just. As David Heyd explains:

> In the present time of entry version, it would make sense for us to avoid savings, irrespective of our actual location in the chain of generations. For regardless of whether our ancestors have saved or not…we have no reason to save for future generations…\(^3\)

The contracting parties—given that they know they will all be members of the same generation—face a familiar type of prisoner’s dilemma: it is collectively in the interests of all generations to cooperate and select certain principles of intergenerational justice (e.g., a just

---

2. Ibid.
savings principle). However, it is in the individual interests of the contracting parties to “defect” and not to select any principles of intergenerational justice. The decision to defect has no effect on the decisions of other generations. These generations either will or will not save. Furthermore, the adoption any intergenerational principles are guaranteed to impose certain costs on the contracting parties.

Rawls is, of course, aware of this problem and attempts to prevent it by grafting on a psychological assumption about the contracting parties. As he explains:

[P]ersons in the original position are not to view themselves as single isolated individuals. To the contrary, they assume that they have interests which they must protect as best they can and that they have ties with certain members of the next generation who will also make similar claims.4

Essentially, rather than modeling the parties as disinterested—as he does in order to determine intragenerational principles of ideal justice—Rawls models the parties as having a concern with at least their immediate descendants. This psychological concern motivates the selection of an intergenerational principle of savings: the contracting parties do not just care about themselves they also care about their immediate descendants; consequently, they select a principle of savings that benefits their descendants.

The addition of the psychological motivation gives rise to a number of objections. Brian Barry, for instance, complains that the psychological amendment is ad hoc:

The only justification offered for the ‘motivational assumption’ is that it enables Rawls to derive obligations to future generations. But surely this is a little too easy, like a conjurer putting a rabbit in a hat, taking it out again and expecting a round of applause.5

---

4 Rawls, A Theory of Justice, 181. A similar claim is repeated later in A Theory of Justice when Rawls considers questions of intergenerational justice in more detail: ‘[T]o achieve a reasonable result, we assume first, that the parties represent family lines, say, who care at least about their more immediate descendants.’ (Ibid. 255).

I think that Barry is completely correct to argue that the assumption seems unmotivated—something simply designed to produce the delivery of the antecedently desired theoretical result. Given Rawls’s commitment to a reflective equilibrium methodology, he is entitled to refine the contractualist framework in order to insure that the results harmonize with certain considered intuitions. However, this process of refinement—as in this attempt to model considerations of justice to future generations by assuming that the parties care about future generations—cannot be so crude that it simply forces the desired output. This conception of the contracting parties should be resisted because it deprives the contractual procedure of any heuristic explanatory value.\(^6\)

I.4 The present time of entry model and an assumption of strict compliance

Rawls, who was characteristically responsive to his critics, abandoned the psychological assumption of concern in his later work but retained his commitment to the present time of entry model. He argued such a model was viable if it was combined with an assumption of strict compliance: the contracting parties—who know that they all belong to the same generation—operate with the knowledge that all actual generations will strictly comply with the principles of justice that they select.\(^7\)

The essential idea is that this assumption of strict compliance is sufficient to overcome the prisoners’ dilemma that I discussed above. The prisoners’ dilemma emerges because members of an individual generation cannot trust the members of other generations to cooperate. Consequently, the dominant strategy is for them to defect and not cooperate, even though it would be in the mutual interest of all generations to cooperate. The assum-

---

\(^6\) For a different objection to Rawls’s position see: Jane English, “Justice between Generations,” *Philosophical Studies* (1977): 91-104. I have chosen not to outline her (good, but more complex) arguments because Barry’s objection is sufficient.

tion of strict compliance, however, means that behind the veil the parties no longer need to worry about trusting the members of other generations to cooperate. They can assume that members of other generations will comply with whatever principles they select. Therefore, given that intergenerational cooperation improves their expected payout, it is rational for them to select principles of intergenerational justice.

1.5 Problems with this appeal to the assumption of strict compliance

This approach initially appears to be promising: it overcomes the prisoner’s dilemma without the addition of an ad hoc motivational assumption. There is, however, a subtle but significant problem with this approach that deprives the contractualist procedure of at least some explanatory heuristic power.

The problem is that Rawls’s model does not elegantly capture the subject matter of intergenerational justice. The subject matter of intergenerational justice concerns what distribution of burdens and benefits between different generations is intergenerationally just. However, given that only members of a single generation are represented in Rawls’s model, it does not represent the distribution of burdens and benefits between members of different generations. Rather, it just represents the interests of a single generation against the backdrop of an assumption that other generations will strictly comply with the principle of justice that they select. Consequently, it models what is just for a single generation to do—ignorant of their relative position in the chain of generations—given that other generations will cooperate. It does not model justice between members of different generations.

In order to feel the force of this problem, it is helpful to reflect on how intuitively problematic it would seem if an analogous model were used to determine ideal principles of intragenerational justice. For instance, suppose Rawls had argued that it was not necessary to include all social positions in the contractual agreement to decide ideal principles of intragenerational justice. Rather, it was only necessary to include a subset of the possible social positions and to generate principles of justice that applied to all social positions against the backdrop of the assumption that the unrepresented parties would strictly comply in the requisite sense. Even if this contractualist model could be used to generate the same set of ideal principles of justice, it would, I submit, be a very implausible approach. This is because it would not explain why the demands of justice have a pressing claim on all of us—
regardless of which social position we actually happen to occupy—because the interests of all the social positions are not appropriately represented in the model.

Ultimately, the problem with this model is that the assumption of strict compliance functions in an implausible sense. As I argued in chapter 2 (§2.5.1), the function of the assumption of strict compliance is to insure that principles of justice are not tainted by the expected non-compliance of actual people. However, in Rawls’s intragenerational contractualist model the interests of all actual people are directly represented because each different social position is included in the model; in contrast, in this intergenerational model the interests of all actual people are not represented—just the interests of a single generation.

I, therefore, conclude that the contractualist procedure should be constructed in such a way that it shows that it is in the rational, individual interest of each and every generation to select principles of justice that are fair to all generations. The key to this—or so I will argue—is to move away from the present time of entry model and to directly represent the interests of each and every generation in the contractualist model.

I.6 The general assembly model

In *A Theory of Justice* Rawls briefly considers the possibility that the parties included in the contractual agreement should consist of a “general assembly” of all actual and possible people from each and every generation. None of these people know which generation they will actually be born into.  

1.6.1 The motivation for the general assembly model

The most obvious motivation for this general assembly model is its naturalness: the task, after all, is to determine principles of intergenerational justice. Consequently, it is appealing to include parties who can represent the interests of each and every generation.

There are also deeper philosophical reasons for adopting this model. Like the present time of entry model and assumption of strict compliance, it is sufficient to overcome the prisoners’ dilemma: none of the parties know which generation they will actually be born

---

into. Consequently, the requisite tension between an individual and group vantage point that generates the prisoners dilemma cannot occur within this model because none of the parties knows which individual vantage point they will in fact occupy. Therefore, if the expected payout for each generation is greater if they cooperate and select mutually beneficial principles of intergenerational justice, then the dominant strategy is for the parties to cooperate.

In contrast to the present time of entry model, the general assembly model appropriately represents the interests of each and every party. It does not represent a subset of interests such as a single generation and assume that other parties—whose interests are not directly represented in the contractual model—will strictly comply with the principles that are selected. Consequently, this contractual model is an explanatorily powerful heuristic: it models obligations and claims between members of different generations; it, thereby, illuminates why all actual people ought to comply with the principles of justice that are selected—because the model impartially captures all of their interests.

1.6.2 Why Rawls rejects the general assembly model

Even though the model seems prima facie plausible it is a model that Rawls categorically rejects. In the original edition of *A Theory of Justice* he states that ‘to conceive of the original position [in this way] is to stretch fantasy too far.’ Interestingly, in the second

Daniel Attas complains that participants in the general assembly model still in fact face exactly the same type of prisoner’s dilemma. He writes: ‘Whatever form an assembly in which all generations attend or are actually represented takes, participants in the assembly know that, though negotiation is possible, the past is fixed and unchangeable. In particular the rate of saving at which previous generations have already saved cannot be altered at this assembly. Hence the rational choice for each generation in the multi-generational assembly interpretation is not to save.’ (Attas, “A Transgenerational Difference Principle,” 196-7.) However, I think that this complaint relies on a confusion: it is true as a matter of fact that the past cannot be changed, however, the meeting is supposed to determine what principles of intergenerational justice it is rational for people to agree to comply with. Consequently, the contractual model is ahistorical and does not have to incorporate the fact that the actual history of the world is in fact fixed. The parties are allowed to adopt the methodological assumption of strict compliance and assume when selecting such principles that actually people will strictly comply with the selected principles.

edition he decides to frame the point slightly differently and writes that it: ‘...would cease to be a natural guide to intuition.’

Rawls does not spell out exactly what he means in either edition. One interpretation, which probably places greater emphasis on Rawls’s remarks in the original edition, is that this model is unacceptable because it diverges too sharply from the way things are in the actual world: as we know in the actual world all generations are not assembled at one time. Rather, each human is mortal and at most—if she is comparatively fortunate—her lifespan overlaps with several other generations. However, if this is the gloss that is put on Rawls’s remarks I do not think his reason for rejecting this model are even prima facie plausible. The aim of a contractual model is to provide a powerful explanatory heuristic; there is no reason to think that such a model could not be powerful in the requisite sense merely because its contemporaneous representation of all generations is different from how things are in the actual world.

An alternative, more plausible, way of interpreting Rawls, which probably places greater emphasis on his remarks in the revised edition, is to see him as making an epistemic point: this version of the contractual agreement is so fantastical that it is not something that we can imagine (at least) sufficiently well. Consequently, it is not something that can serve as a powerful explanatory heuristic because it is not something that we can imagine sufficiently well enough to consider what principles the contracting parties would ascent to.

1.6.3 Refining the general assembly model into the representation of all generations model

For the sake of argument, I grant that this epistemic objection to the general assembly model is successful. However, I will show that the general assembly can be refined into what I term the representation of all generations model. This epistemic objection does not apply to this refined version of the model.

In the context of an intragenerational contractualist model, it has become conventional to refer to multiple parties. However, as Jean Hamilton observes ‘...there is no

---


12 Interestingly, even in the original edition Rawls hints that he may mean something like this when he goes on to explain: ‘In any case, it is important the original position be interpreted so that one can at any time adopts its perspective.’ (Rawls, *A Theory of Justice: Original Edition*, 139.)
theoretical reason to posit more than one party….”¹³ This is because given that all the parties would be situated in exactly the same way a single party would be sufficient.¹⁴ Similarly, in an intergenerational context a single party ignorant of which generation they will actually occupy would also be sufficient because all the parties would be identically situated. Consequently, a general assembly of all possible and actual people from each and every generation is not necessary.¹⁵


¹⁴ Samuel Freeman attempts to push back on the claim that a single party would be sufficient. He writes that: ‘One reason Rawls insists on an agreement [between multiple parties] is to carry through the contrast with utilitarianism…The original position is not set up like Ideal Observer theories, where a single judge with full knowledge of all but its own identity impartially represents and sympathetically identifies with the interests of everyone…Were there but one party choosing principles, there would be no one for him to commit himself to. Agreement, unlike individual choice, implies a joint undertaking where each is held by the others to his decision, thereby ensuring the perpetuity and irrevocability of the principles agreed to… Even if the parties are symmetrically situated by that condition, Rawls needs to maintain the idea of distinct individuals coming to an agreement to make the main idea of his theory go through.’ (Samuel Freeman, “Reason and Agreement in Social Contract View,” in *Justice and the Social Contract: Essays on Rawlsian Political Philosophy* (New York: Oxford University Press, 2006), 17-44, p. 35) It is not clear to me that Freeman’s argument succeeds: the idea of different parties coming together certainly has rhetorical force; however, this does not mean that it performs an essential role in Rawls’s actual philosophical argument. However, even if Freeman’s argument were to succeed, it would only be sufficient to show that in both an intragenerational and intergenerational context two parties were necessary.

¹⁵ Arguably there are additional reasons to refine the general assembly model into the representation of all generations model. Heyd argues that the general assembly model would be logically incoherent because we cannot coherently conceive of a contractualist procedure between parties whose very existence is at stake in the negotiations. Their very existence is at stake, Heyd argues, because considerations of intergenerational justice are inseparable from those of population ethics. Essentially, the idea seems to be that the determination of principles of intergenerational justice will have a direct effect on the size of the population. Therefore, in selecting principles of intergenerational justice the contracting parties must determine which contracting parties actually get to exist. However, Heyd argues that this is logically absurd because ‘[a]n assembly of individuals cannot decide on its own size.’ (Heyd, “A value or an Obligation,” 173.) (Attas raises similar concerns: see Attas, “A Transgenerational Difference Principle,” 196.) However, on the representation of all generations model there is just a single party; consequently, the selection of principles does not have an effect on the existence of the single party even if the selection of such principles has implications for the actual population.
It is important to clarify that even if there is just a single party the interests of each and every generation can be *directly* represented in the contractualist model. They are represented in the model because the single party is ignorant of which generation they will actually be born to and, consequently, it is rational for the party to select a principle that is intergenerationally impartial—hence fair—in the requisite sense. Note, in contrast to Rawls’s claim that the contracting parties know they all belong to a single generation the single contracting party does not know when she will be born—hence she selects principles of justice that are intergenerationally impartial between members of different generations, rather than rational for members of her own generation in light of the fact that other generations will strictly comply with the principles she selects.

Given that in an intergenerational context no further parties are required—just an additional assumption of ignorance about which generation the party is born—it is difficult to see how this could offer any less of a guide to intuition that in the intragenerational case. Consequently, the epistemic objection is overcome, or at the very minimum does not apply with any more force to the intergenerational contractualist procedure than in the intragenerational contractualist procedure.

1.7 Two objections and responses

Suppose that I am right to contend that the representation of all generations model compares favorably with other contractualist models of intergenerational justice. Even still, there are two challenges to the very possibility of developing a theory of intergenerational justice within a Rawlsian contractualist framework that have emerged in the literature. I end this appendix by outlining and responding to these challenges.

1.7.1 Maximin and the first generation

The first challenge recently defended by Stephen Gardiner is that, given the choice problem posed, it will be rational for the contracting parties (or strictly speaking the single


16 For further discussion, see: Attas, “A Transgenerational Difference Principle.”
contracting party) not to select any intergenerational principles of justice that benefit future generations. This is because such sacrifices would benefit every generation except for the first generation. Maximin reasoning directs the contracting parties to favor the least advantaged (i.e., in this case the first generation). Therefore, these parties should not adopt any principles of intergenerational principles of justice.\textsuperscript{17}

It is worth pausing to note that in the context of nonideal theory this would clearly be a disastrous result; it would entirely block the intergenerational transition to ideal justice: given that maximin reasoning requires prioritizing the interests of the least advantaged—in this instance the first generation no principles of justice that impose burdens on all generations for the sake of future generations, including the first, would be selected.

Daniel Attas suggests that the idea of a first generation involves a theological idea of creation more suited to mythology than philosophy. Consequently, it has not relevance for us in the actual world.\textsuperscript{18} However, although the idea of a first generation of humanity seems mythical, the idea of the first generation of a newly founded country, such as Pakistan, does not. A better response is that the claim that later generations of a society are always wealthier than the first generation of a society is false. This is because, for instance, if the first generation were to squander all of their natural resources for their own short-term gain, later generations would almost certainly be poorer. Therefore, given that actual empirical conditions are completely uncertain behind the veil of ignorance, the contracting parties would not reason from the contingent empirical assumption that the wealth of generations will inevitably grow over time. Consequently, maximin reasoning does not block the adoption of intergenerational principles of justice that benefit later generations.

I.7.2 Heyd’s skeptical challenge

Heyd argues that if we accept Rawls’s conception of the circumstances of justice it is not possible to offer an intergenerational theory of justice. His argument can be reconstructed as follows:

\textsuperscript{17} See: Gardiner, “A Contract on Future Generations,” 110. (Rawls himself notes in passing that, ‘[I]t is immediately obvious that every generation, except possibly the first, gains when a reasonable rate of saving is maintained.’ (Rawls, \textit{A Theory of Justice}, 256.))

Premise 1: Justice applies if and only if the circumstances of justice obtain.

Premise 2: ‘[H]uman cooperation is possible and necessary only when human beings who live side by side on the same territory are concerned… The circumstances of justice are thus essentially associated with the conditions of mutuality or reciprocity… [J]ust cooperation takes place only where human beings are mutually vulnerable as well as capable of benefiting each other.’

Premise 3: ‘[O]n the intergenerational dimension dependence seems to be in principle unidirectional, i.e. involves relations which are not and will never be based on reciprocity. We are dependent on the behavior of our ancestors but they are in no way dependent on ours.’ Consequently, the circumstances of justice do not obtain between different generations, which do not actually overlap.

Conclusion: It is not possible to give an intergenerational theory of justice (at least between generations that do not actually overlap with one another), because the circumstances of justice do not obtain.

The argument is clearly valid. Premise 1 is analytic; premise 3 is a clear and uncontroversial empirical fact.

An obvious limitation of the argument is that it does not show that intergenerational justice is impossible, just that it is impossible between generations that do not actually overlap with one another. As David Gauthier writes:

The generations of mankind do not march on and off the stage of life in a body, with but one generation on stage at a time. Each person interacts with others both older and younger than himself, and enters thereby into a continuous thread of interactions…

---

20 Ibid. 169.
The fact that each generation overlaps (at least standardly) with other generations may be sufficient to establish quite robust principles of intergenerational justice. For instance, each generation might have a duty to save for the next generation, who in turn have a duty to save for the next generation etc. However, in a nonideal context Heyd’s argument has bite. This is because, as I argued in chapter 2, in such a context, different generations may have duties to future generations (e.g., to help promote their ability to exercise their basic liberties) that they do not actually overlap with and, consequently, cannot reciprocally cooperate with.

The argument can be overcome by rejecting premise 2. The problem with the premise is that Heyd interprets the crucial concepts of “possibility” and “necessity” in a way that a Rawlsian can coherently resist. Heyd construes the requisite “possibility” as referring to the abilities of actual people; “necessity” as referring to what it is rational for actual agents to do in order to mutually accrue benefits and to avoid relations of mutual vulnerability.

Heyd is clearly correct that it is not possible for actual people to cooperate (if cooperation is understood reciprocally) in an intergenerational context, however, this does not entail that Rawlsians cannot offer a contactualist theory of intergenerational justice. This is because a contractual theory of intergenerational justice can be constructed which models the contracting parties in such a way that it is possible and rational for them to intergenerationally cooperate even if, as a matter of fact, it is not possible for actual people to intergenerationally reciprocally cooperate. The contractualist model, essentially, does not have to necessarily reflect the predicament of actual people—it just has to provide an explanatory heuristic in order to illuminate what justice requires.22

In order to clarify why Rawlsians can reject the concept of “necessity” that Heyd’s argument relies on, it is illuminating to compare Rawls’s position with that of David Gauthier. The contractarian theory of justice that Gauthier develops in *Morals by Agreement* is self-consciously situated in the Hobbesian tradition. As Samuel Freeman explains, ‘In Hobbesian

---

22 Similar claims are made by Daniel Attas, who writes: ‘Still, this [i.e., the fact that in an intergenerational context it is impossible to engage in a mutually cooperative endeavor] is a feature of the real world, but it need not be the situation in the original position. If it is required to extend the scope of justice so that it covers relations between generations, then the original position can be reconstructed so as to model relations of mutuality between generations even if these are absent from the real world.’ (Attas, “A Transgenerational Difference Principle,” 191.)
views, cooperation for mutual advantage involves no irreducibly moral elements. Hobbesian views aim to show that morality is a subordinate notion, grounded in individuals’ antecedent desires and interests.\(^ {23} \)

Given Gauthier’s Hobbesian commitments, it would clearly be impermissible for him to endorse a contractualist procedure in which the parties were modeled in a way that was different from what it was individually rational for actual agents to do in order to maximize their own interests. This is because, for Gauthier, the aim of moral and political philosophy is to respond to the moral skeptic by showing that an amoral or non-cooperative person fares less well then a compliant agent in a cooperative context and, therefore, it is individually rational for agents to comply with certain moral principles.\(^ {24} \)

However, in contrast, given that Rawls is embedded in a Rousseauean tradition in which principles of right and justice cannot be accounted for without appealing to certain irreducible moral notions,\(^ {25} \) it does not matter that the contractualist model diverges from a direct representation of what it is rational for actual individual agents to agree to in order to maximize their own individual interests. The contractualist framework can be constructed in such a way that it embeds irreducible moral considerations and, therefore, insures that the rational decision that the contracting parties arrive at reflects considerations of justice, rather than what it is rational for actual, individual agents to agree on.

Although this option is available to Rawlsians, it is conspicuous that Rawls himself often seems torn on this issue. Indeed, it appears as if he never resolved it to his own satisfaction. As he explains in *Political Liberalism*, “reciprocity”—a central feature of the circumstances of justice:


\(^{24}\) Indeed, when Gauthier, briefly, gets around to discussing intergenerational justice the argument that he makes for a limited set of intergenerational obligations exclusively hinges around an appeal to individual agents’ interests. As he explains: ‘Each person, in considering the terms on which he is to cooperate with those in an earlier generation than himself, must keep in mind his need to establish similar terms with those of a later generation, who in turn must keep in mind their need to cooperate with members of a yet later generation, and so on.’ (Gauthier, *Morals by Agreement*, 299). (See also: David Gauthier, “Bargaining and Justice” in *Ethics and Economics*, ed. Ellen Paul & Fred D. Miller (Oxford: Blackwell, 1985), 40-45.)

‘[L]ies between the idea of impartiality, which is altruistic (being moved by the general good), and the idea of mutual advantage understood as everyone’s being advantaged with respect to each person’s present or expected future situation as things are.’

A careful reading of Rawls reveals that he sometimes explicates the concept of reciprocity in different ways. For instance, at the beginning of *Political Liberalism* he writes:

For these terms to be fair terms, citizens offering them must reasonably think that those citizens to whom such terms are offered might also reasonably accept them… And they must be able to do this as free and equal, and not as dominated or manipulated, or under the pressure of an inferior political or social position. I refer to this as the criterion of reciprocity.

However, later in the book he writes:

Fair terms of cooperation specify an idea of reciprocity: all who are engaged in cooperation and who do their part as the rules and procedure require, are to benefit in an appropriate way as assessed by a suitable benchmark of comparison. A conception of political justice characterizes the fair terms of cooperation.

This later explication suggests that justice governs the benefits and burdens of *actual* social cooperation. The former explication suggests that reasonable acceptability is sufficient; there is no reference to actual cooperation.

I suggest that the former explication is more plausible. This is particularly the case in the context of nonideal theory. If the demands of nonideal justice could be completely captured by an intergenerational rate of savings principle then perhaps the possibility of actual social cooperation could ground the applicability of justice. However, it does not.

---

27 Ibid. Xliv.
28 Ibid. 16.
Such demands of justice also include duties to distant generations—such as the duty to promote future generations ability to be able to exercise their basic liberties. Consequently, the concept of actual social cooperation simply does not seem rich enough to ground the requisite normative account of what should be done in nonideal conditions. The normative demands of this account seem to transcend relations of actual social cooperation.

I.8 Conclusion

In this appendix I have outlined and defended the representation of all generations model that I use in chapter 2. I argue that this model compares favorable with other prominent contractualist models of intergenerational justice. Given the focus of my inquiries I defend this intergenerational contractualist model in order to generate nonideal principles of justice. However, I also think that it is an appropriate framework to generate ideal intergenerational principles of justice; in particular, it can be used to assess Rawls’s claim that the only intergenerational principle of justice is a sufficientarian principle of just savings that in real terms may fall to zero after ideal institutions have been established.29

Bibliography


Butt, Daniel. *Rectifying International Injustice. Principles of Compensation and Restitution Between Na-


Kingkade, Tyler. “Pat McCrory Lashes out Against ‘Educational Elite’ and Liberal Arts College Courses,” *Huffington Post* (2013), New York: http://www.huffingtonpost.com/2013/02/03/pat-mccrory-college_n_2600579.html


Korsgaard, Christine M. *Creating the Kingdom of Ends*. Cambridge: Cambridge University Press, 1996.


Sandel, Michael J. Liberalism and the Limits of Justice. Cambridge: Cambridge University Press,
1998.


