

THE FRAGILITY OF JUSTICE

Political Liberalism and the Problem of Stability

Jeffrey William Howard
Nuffield College, Oxford
April 2013

Thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Politics

in the Department of Politics and International Relations at the University of Oxford

98,967 words

Acknowledgements

This doctoral dissertation grew out of a series of essays about how political liberals should think about and respond to wrongdoing—be it the wrongdoing perpetrated by vicious criminals, self-serving or fearful majorities, or citizens opposed to the fundamental values of liberal democratic civilisation. In time it grew into a sustained reflection on a duty at the heart of our political and moral life: the duty to maintain the smooth functioning of the very capacity that makes us moral agents in the first place.

I would not have been able to complete this journey without the assistance of many people. I am thankful to the members of the Nuffield Political Theory Workshop for having read and discussed my work on four distinct occasions, enabling me to identify and craft responses to various objections. I thank Neal Carrier, Ian Carroll, and Caleb Yong for offering wonderful suggestions on many chapters. I extend especial gratitude to Malte Frøslee Ibsen for probably hundreds of hours of discussion on Rawls and contemporary political philosophy. He has often helped me to realise what I was trying to argue before I realised it myself.

This task would have been impossible without my three mentors, who have taught me—in their different ways—how to do normative political philosophy. David Miller has been indispensable throughout the whole process, both in supervising my MPhil and subsequently serving as my college adviser. His wise comments on the penultimate draft of this thesis helped me considerably in the final stages of revision. Bob Goodin has offered numerous challenges to my arguments over the past year, resulting in many improvements. Finally, I thank my supervisor, Jeremy Waldron. From the moment I spotted him sitting behind me on a British Airways flight between Pisa and London—prompting me to introduce myself at 30,000 feet and entreat him to supervise my DPhil, his alarmed eyes searching nervously for the emergency exit—he has been a constant, energetic source of wisdom and support. Without those countless sessions thinking through my ideas over lattes in All Souls, I would still be wandering in the theoretical wilderness.

Most of my thanks goes to my parents, for teaching me to care about—and not give up on—the value of others, which places demands on us that I am only beginning to understand; and to Josie Delap, who discussed every element of the argument with me from its infancy and scrutinised the final draft with meticulous care. She fortifies me.

ABSTRACT

The Fragility of Justice

Human powers of moral reasoning and motivation are fragile. How should citizens committed to the achievement of liberal justice respond to this fact? This dissertation theorises a class of moral requirements that are central to the practice of liberal democracy but have been recently overlooked by political philosophers: *the fortificational duties*, which enjoin citizens to design and submit to civic practices that improve both their moral reasoning and the motivational resilience of their sense of justice. It considers the proposition that a conception of justice is unjustified if unlikely to generate its own freely willed maintenance, or *stability*, in the face of human nature, and it argues that this proposition is false. If justice may face overwhelming resistance unless steps are taken to fortify ourselves against our own fallibility, the right response is to pursue precisely such fortification.

Chapter One sketches the orienting ideal of the dissertation: an ideal of a social world in which citizens live together as free and equal. Chapter Two assesses the proposition that we ought to modify or abandon this ideal if we determine that it is unlikely to be freely realised without serious civic or institutional assistance—a move suggested by John Rawls’s “stability test”—and it argues that the candidate arguments for this conclusion fail. The chapter instead argues that citizens are subject to moral requirements to fortify their sense of justice by designing and submitting to measures that increase the likelihood that they will accurately identify and freely comply with their fundamental moral duties. These measures together constitute a liberal democracy’s “stability charter.”

Chapters Three to Six explore proposed elements of citizens’ stability charter. Chapter Three discusses the fortification of moral reasoning through democratic deliberation. Chapter Four considers what institutional mechanisms could keep democracy oriented toward the achievement of justice despite human fallibility, and it defends a minimalist conception of judicial review as a case study. Chapter Five argues that the practice of criminal punishment is justified by the duties of wrongdoers to pursue additional fortificational measures in the aftermath of their moral powers’ defective operation. And Chapter Six focuses on the special problem posed to the enduring achievement of justice by “unreasonable citizens” who reject fundamental liberal values.

The distinctive contribution of the dissertation lies, firstly, in its novel appropriation of the Rawlsian ideal of stability—reconceiving stability not as a justificatory condition set by reason on our convictions, but as a practical challenge that our own convictions set *for us*—and, secondly, in its deployment of that insight to motivate novel arguments about the character of democratic deliberation, the limits and role of judicial review, the proper purposes of criminal punishment, and the ideal method of engagement with unreasonable citizens.

CONTENTS

<i>Introduction</i>	p. 1
Chapter 1. <i>Liberal Foundations</i>	p. 15
Chapter 2. <i>The Problem of Stability</i>	p. 44
Chapter 3. <i>Democratic Deliberation and the Fortification of Moral Reasoning</i>	p. 111
Chapter 4. <i>Democratic Stability and Judicial Review</i>	p. 143
Chapter 5. <i>The Fortificationist Theory of Punishment</i>	p. 174
Chapter 6. <i>Deliberative Assimilation</i>	p. 230
<i>Conclusion</i>	p. 274

“[T]he problems of political philosophy...are practical all the way down.”¹

– Christine Korsgaard

“If justice perishes, then it is no longer worthwhile for men to live upon the earth.”²

– Immanuel Kant

¹ Christine Korsgaard, “Realism and Constructivism in Twentieth Century Moral Philosophy,” in *The Constitution of Agency: Essays on Practical Reason and Moral Psychology* (New York: Oxford University Press, 2008), p. 318.

² This unattributed translation (perhaps Rawls’s own) appears in John Rawls, *The Law of Peoples* (Cambridge: Harvard University Press, 1999), p. 128n. An aesthetically inferior translation can be found in Immanuel Kant, “The Metaphysics of Morals,” in *Practical Philosophy*, edited and translated by Mary Gregor (New York: Cambridge University Press, 1996), p. 473.

INTRODUCTION

The Fragility of Justice

I

HUMAN BEINGS ARE DANGEROUS. We are capable of unleashing unfathomable damage upon one another's interests. And we do: murder, rape, torture, religious persecution, political oppression, ethnic cleansing, racial apartheid, starvation, humiliation, trafficking, enslavement—the list of what we have done, and continue to do, stretches too long. Behind these evils resides a diverse range of responsible vices: cruelty, greed, jealousy, thoughtlessness, malice, arrogance, ignorance, vengefulness, apathy, and unreflective fanaticism. Philosophers continue to wonder whether these vices belong, in some sense, to our nature, or whether, as Rousseau believed, they are triggered by morally unjust social arrangements.¹ It is true that violence has declined over time²—though, as Michel Foucault would have insisted, we cannot infer the absence of inequalitarian subordination from a lack of easily detectable brutality.³ Nor can we test Rousseau's hypothesis that the worst evils would disappear within perfectly just societies, for there are none to be found. Instead we must simply live with these evils, and contemplate what to do about the dangerous species that engenders them.

What shall we do with ourselves? Kant believed that if human beings were incapable of justice, life would not be worth living.⁴ If we are inclined to join Kant's horror at the prospect of a world in which moral convictions remain impotent and forever aspirational, and in which terror and conflict rule, then the question of whether justice is compatible with our nature is one with high stakes indeed. But the grim portrait

¹ Jean-Jacques Rousseau, "Discourse on the Origin of Inequality," in *The Basic Political Writings*, translated by Donald A. Cress (Indianapolis: Hackett Publishing Company, 1987).

² Steven Pinker, *The Better Angels of Our Nature* (London: Penguin Books, 2011).

³ Michel Foucault, *Discipline and Punish* (London: Penguin Books, 1991).

⁴ Kant, "The Metaphysics of Morals," p. 273.

of humanity's prospects that is suggested by the third-person perspective is strikingly resisted by our first-person experience. From the perspective of an agent—the perspective of you or me—the *possibility* of justice seems obvious. It is something we assume, inescapably, every time we choose not to harm a rival; every time we choose to offer assistance to a stranger; every time we vote in support of a new policy that secures rights to a previously disenfranchised minority. One of my central claims will be that our sense of ourselves as persons capable of seeing the force of moral reasons for action, and choosing to regulate our conduct in accordance with those reasons, does not simply *testify* to the possibility of justice; it *entails*, on pain of incoherence, the possibility of justice. If the possibility of justice is not real, it seems, then neither are we.

What explains the incongruence between the third-person view of humanity and our experience as agents? One explanation is that our self-understanding is illusory. This is a worry Kant took seriously. We *feel* as though we are capable of justice, and we presuppose that capacity in all we do—but we are mistaken. Kant registers his fear:

If it is a sight worthy of a divinity to see a virtuous man struggling with adversity and temptations to evil and yet holding out against them, it is a sight most unworthy, I shall not say of a divinity but even of the most common but well-disposed human being to see the human race from period to period taking steps upward toward virtue and soon falling back just as deeply into vice and misery. To watch this tragedy for a while might be moving and instructive, but the curtain must eventually fall. For in the long run it turns into a farce; and even if the actors do not tire of it, because they are fools, the spectator does, when one or another act gives him sufficient grounds for gathering that the never-ending piece is forever the same.⁵

The proposition that justice is beyond the bounds of the possible is disorienting. Yet if it is false, it is also the most pernicious proposition there ever was. It releases those in power from having to do anything. Were humanity's quest for justice to be halted in the name of impossibility, countless lives would be blighted, projects dashed, relationships

⁵ Immanuel Kant, "On the Common Saying: That may be correct in theory, but it is of no use in practice," in *Practical Philosophy*, pp. 305-6.

destroyed. For many of us, the stakes are too high—and the decks of self-interest too suspiciously stacked—for such complacent assumptions to be warranted, especially when they force incomprehensible revisions to our self-understanding. This brings us to a second, less paralysing explanation of the lack of congruence we encounter: justice remains absent from the world because we have simply not worked hard enough to bring it to the world. We have not adequately reckoned with our own fallibility to mitigate the elements of our nature that subvert sound moral reasoning and decision-making. Rather than suicide or moral resignation, this explanation prompts a positive response, one I shall endorse: *get to work*.

There are two practical challenges that confront us: one concerning how to *transition* to a just world from our presently unjust perch; another concerning how people might *sustain* a just world once it is up and running. For reasons I shall explain, I believe the latter is the philosophically prior project, and it is this project to which this dissertation contributes. Suppose, fancifully, we awake one morning to find ourselves living in a just society—or, taking into consideration the perennial fact of reasonable disagreement about justice, living in what I shall call a *fundamentally* just society. We have been socialised in such a way that we have acquired *a sense of justice*, a disposition to act justly toward others; we have been educated to hold reasonable (i.e., liberal) convictions about its demands; and we understand that it is in the nature of our moral convictions that they are *regulative* over our practical reasoning, setting the limits of everything that we may do. We share what I will call *a commitment to justice*—a commitment to a social world in which we live together as free and equal by regulating our social relations by principles that are mutually justifiable. We live in (a variant of) what John Rawls calls a *well-ordered*

society—a society in which our shared commitment to justice is publicly realised in our relations.⁶

But all is not well. We sense the fragility of our civic practices, and we locate that fragility in two sources within our nature. Firstly, we know ourselves to be, for lack of a better term, *epistemically fallible*. Our powers of moral reasoning are susceptible to mistakes—as evidenced by the prodigious quantity of moral error in human history (including error on how to understand the very ideals of freedom and equality) and overwhelming social-scientific evidence of the imperfections of human cognition. Secondly, our powers of moral motivation are susceptible to subversion from manifold countervailing forces; we are *motivationally fallible*. Even if our wider doctrines about the meaning and nature of human life and the good are compatible with our convictions about justice—even if “the good” and “the right” are congruent within our belief-set⁷—we nevertheless still face temptations to acquire more than what justice permits us, as well as emotional impulses to violate others’ rights when they anger or upset us. Taken together, this *fact of fallibility* raises a sobering question that the reasonable citizens of a fundamentally just, well-ordered liberal democracy ask themselves. While it may be true that justice *could* endure, we must nevertheless ask: *will it* endure? Is our capacity for justice strong *enough*? Or is our nature likely to operate as a slow-acting poison on just institutions, a corrosive acid that eats through reasonableness to bring our civic life to a state of moral decay?

⁶ John Rawls, *A Theory of Justice*, revised edition (Cambridge: Harvard University Press, 1999), p. 4.

⁷ See Rawls, *A Theory of Justice*, pp. 450ff.

This is an old question. It concerned Plato in *The Republic* when he assessed the prospects of degeneration within the ideal city;⁸ it concerned Thomas Hobbes in *Leviathan* when he considered the forces that trigger a commonwealth's dissolution;⁹ and it concerned John Rawls in his own *magnum opus* when he asked whether a liberal conception of justice would be *stable*.¹⁰ That question dominated the third and final section of *A Theory of Justice*, and Rawls's dissatisfaction with the answer he provided motivated radical changes between that book and his later work, *Political Liberalism*.¹¹ In *A Theory of Justice*, Rawls optimistically avowed that his preferred conception of justice, *justice as fairness*, would arouse a strong sense of justice in the members of any society regulated by it. He also argued that citizens' sense of justice would be supremely regulative, and precisely because they would find that the demands of justice rested in harmonious concord with their own good. And they would find this precisely because members of a just society could be expected to share a partial conception of the good according to which they would have an intense desire to express their nature as autonomous beings—precisely the nature that acting justly, on a Kantian view, affirms. Rawls came to see, however, that citizens would not share a conception of the good life in a society regulated by liberal principles, after all. For in such a society citizens' free use of reason—encouraged by their substantively protected liberties of conscience, expression, and association—would lead to a *diversity* of views on matters of the good. Hinging the case for a conception's stability on adherence to a single shared conception of the good,

⁸ Plato, *The Republic*, second edition, translated by Allan Bloom (New York: Basic Books, 1968), pp. 221ff.

⁹ Thomas Hobbes, *Leviathan*, edited by C. B. Macpherson (London: Penguin Books, 1968), pp. 363ff. And also: "And though Sovereignty, in the intention of them that make it, be immortal; yet is it in its own nature, not only subject to violent death, by foreign war; but also through the ignorance and passions of men, it hath in it, from the very institution, many seeds of a naturall mortality, by Intestine Discord" (p. 272).

¹⁰ John Rawls, *A Theory of Justice*, pp. 347-514.

¹¹ John Rawls, *Political Liberalism* (New York: Columbia University Press, 2005).

then, was implausible. Abandoning this strategy, Rawls instead imagined that an *overlapping consensus* would be the stabilising mark of the well-ordered society, with citizens endorsing liberal justice through their own distinct paths. In this way, the stability of justice could be achieved despite ethical diversity. Let us call the vision of how an ideal of justice is to be freely realised its *stability charter*.

It is important to see what *kind* of answer Rawls provided to the question of whether a well-ordered society would backslide into justice. His stability charter was not a user's manual for citizens—an account of the steps they might take to render enduring their vision of a just order. Rather, Rawls's purposes resembled Rousseau's in his *Discourse on the Origin of Inequality*.¹² In that work, Rousseau aspired to meet the sceptic's challenge. Hobbes believed that human nature was too fragile for peace to be freely willed—that coercion was necessary to keep our fear and violent impulses at bay.¹³ Rousseau aspired to give people hope in the possibility of a future where justice *could* be freely willed. He could provide people with that hope if he could give them a plausible story about how the elements of their nature that *seemed* evil had really been introduced by malfunctioning social institutions. Rawls, too, aspired to give hope to humanity; he set himself the sceptic's challenge, and he aspired to meet it. In so doing, Rawls's aim was not, of course, to offer hope to the imaginary citizens of some well-ordered society far in the future. It was to offer hope to us.

I do not wish to denigrate the motivational significance of hope. But I cannot help but wonder whether the concerned citizens of the well-ordered society that I imagined a moment ago—nervous about the prospects of their fragile relations' enduring maintenance—would regard their worries as allayed by the *kind* of response that

¹² Jean-Jacques Rousseau, "Discourse on the Origin of Inequality," in *The Basic Political Writings*, translated by Donald A. Cress (Indianapolis: Hackett Publishing Company, 1987), pp. 39, 60-81.

¹³ Hobbes, *Leviathan*, pp. 183ff.

Rousseau and Rawls provide as to why their just arrangements are unlikely to backslide into injustice. Agents who hold a commitment to justice but are concerned that their fallibility will imperil that commitment's realisation do not simply want hope. They want to know what to do. While it is artificial to ask ourselves what imaginary agents might agree to do in a time and a place that have never been, we are not making a prediction; we are investigating what it would be reasonable for such persons to think. Given all we know about human epistemic and motivational fallibility, I do not believe that reasonable citizens committed to one another's status as free and equal would be content to sit idly by and hope that justice sustains itself. If our partly idealised citizens should be dissatisfied by Rawls's promise—if they should demand more from their political philosophers—then so should we.

This dissertation reconceives what a liberal democracy's stability charter ought to be. My general thesis, simply stated, is that a serious commitment to justice involves working hard to determine how just relationships among citizens can be realised in the face of obstacles to their achievement, including obstacles generated by so-called human nature. I shall argue that citizens are subject to fortificational duties to design and submit to civic practices that improve both their moral reasoning and the motivational resilience of their sense of justice, and that many important elements of our civic life are best conceived and justified as part of citizens' shared strategy for making justice a reality. These propositions, if true, account compellingly for why engagement with the world of "real politics" is necessary for political theory, as the "realist" movement has insisted,¹⁴ but they do so in a manner that is morally uncompromising and thereby congenial to the

¹⁴ For example: Raymond Geuss, *Philosophy and Real Politics* (Princeton: Princeton University Press, 2008); Bonnie Honig, *Political Theory and the Displacement of Politics* (Ithaca: Cornell University Press, 1993); Bernard Williams, *In the Beginning Was the Deed* (Princeton: Princeton University Press, 2007);

very philosophers such realistic critics have targeted:¹⁵ it is we who must bend our practices to suit the demands of justice, not the other way around.¹⁶

II

A dissertation on the stability of our capacity for justice naturally begins with some account of what “justice” consists in. It could be otherwise. Any account of morality that aspires to guide agents’ deliberation—to illuminate reasons for action—will, I believe, need a supplemental account of stability: of how those agents are going to come to believe in those reasons and regulate their conduct effectively by them. The project I undertake here could thereby proceed in a generalised form, sketching the general framework of stability that *any* moral doctrine has reason to embrace, regardless of what the doctrine’s avowed moral standards might be. Utilitarians and theocrats, if they care about the realisation of their ideals, should be as interested in stability as much as political liberals should.

Why undertake a case study on *political liberalism* in particular? Firstly, Rawlsian contractualism remains the central game in town in contemporary Anglo-American political theory. Robert Nozick’s declaration that “[p]olitical philosophers now must either work within Rawls’ theory or explain why not”¹⁷ remains largely true, as the raging debate of the day—on global justice—makes manifest.¹⁸ Secondly, Rawlsian

¹⁵ The characteristic target of the realists is John Rawls. But a purer target is G. A. Cohen, *Rescuing Justice and Equality* (Cambridge: Harvard University Press, 2008).

¹⁶ I take the language of “bending” from David Estlund, “Human Nature and the Limits (If Any) of Political Philosophy,” *Philosophy & Public Affairs* 39, 3 (2011) 207-237, p. 237.

¹⁷ Robert Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974), p. 183.

¹⁸ Many are convinced that the truth about global political morality hinges on how we understand what Rawls means by “the basic structure.” For a sophisticated analysis of the recent literature with this character, see Arash Abizadeh, “Cooperation, Pervasive Impact, and Coercion: On the Scope (not Site) of Distributive Justice,” *Philosophy & Public Affairs* 35, 4 (2007): 318-58. The most significant early attempt to use the Rawlsian framework for purposes of global political theory

contractualism is a considerably developed doctrine that has grappled extensively with the question of stability. However, because I think it misunderstands the nature of stability's normative significance, stability has not secured the place it deserves on political philosophy's research agenda. This dissertation project serves as crucial preparation to future work to develop precisely that agenda. Finally, I set my gaze on Rawlsian contractualism because I contend that its approach to questions of political morality is both arresting and superior to rival views; its methods and conclusions are, I believe, essentially correct. I do not argue for that grand thesis in this dissertation. But convictions guide action, including the selection of intellectual projects, as Max Weber long ago observed.¹⁹ We should not give the theocrats an account of how to stabilise their preferred conception of justice. On this matter, I affirm the view of thinkers as diverse as Ronald Dworkin and Max Horkheimer: academic political philosophy, at its best, does not stand *apart from* or *above* first-order debates about what we ought to do in politics.²⁰ Rather, it is *part* of politics. Many will doubt that proposition.²¹ But while the project is motivated by this idea, it does not depend on it. For now, then, I shall simply

was Charles Beitz, *Political Theory and International Relation* (Princeton: Princeton University Press, 1979), pp. 127ff.

¹⁹ Max Weber, "Science as a Vocation," in *From Max Weber: Essays in Sociology*, edited and translated by H. H. Gerth and C. Wright Mills (New York: Oxford University Press, 1946), pp. 145ff.

²⁰ Ronald Dworkin discusses this theme in many places, but for one example, see "Hart's Postscript and the Point of Political Philosophy," in *Justice in Robes* (Cambridge: Harvard University Press, 2006), pp. 140ff. Horkheimer says that "materialism" (his term at the time for critical theory, or—for our purposes—political philosophy) "understands itself as the theoretical side of the effort to rid the world of its existing misery." Max Horkheimer, "Materialismus und Moral," *Gesammelte Schriften, Band 3, Schriften 1931-36* (Frankfurt: Fischer Verlag, 2009), p. 131. I am grateful to Malte Frøslee Ibsen for the translation from the German.

²¹ Because this view is oriented to determining and acting upon our *justified* reasons for action, it is not incompatible with the view expressed in Daniel McDermott, "Analytical Political Philosophy," in *Political Theory: Methods and Approaches*, edited by David Leopold and Marc Stears (Oxford: Oxford University Press, 2008), p. 25. Cf. Andrew Rehfeld, "Offensive Political Theory," *Perspectives on Politics* 8, 2 (2010): 465-86.

follow Joshua Cohen: “In saying that political philosophy is a practical subject, I am not offering a definition or an argument but simply declaring a position.”²²

Thus the dissertation’s arguments unfold within the contractualist framework of Rawlsian political liberalism. That formidable label already indicates the general character of the view: it is *liberal* in that it is committed to universal (i.e., equal) freedom for all; it is *politically* liberal in that it grounds and elaborates that aspiration through principles about right and wrong in the social and political domain, not a comprehensive theory of meaning, truth, and goodness in all spheres of human life; and it is *contractualist* in that it conceives the specific norms of conduct demanded by this commitment as the conclusions of an idealised reasoning process that is fruitfully conceived as some kind of hypothetical social contract among faultlessly reasoning agents.

What is it to hold a commitment to justice according to this framework? Chapter One sketches the fundamental elements of this question’s answer. To have a commitment to justice, on this view, is, firstly, to view others in accordance with a particular understanding of human moral personality that regards persons as *free and equal*—as moral agents of equal worth who ought to be respected as capable of framing, revising, and pursuing their chosen conception of a good life; secondly, to identify terms of social cooperation that are mutually acceptable among persons so understood; thirdly, to attempt to resolve and provisionally settle reasonable disagreements about what those terms are *democratically*; and fourthly, effectively to regulate one’s relations with others by those terms.

Chapter Two turns to the central problem of the dissertation. I will consider John Rawls’s thesis that our ideal public philosophy is only justified if it would generate

²² Joshua Cohen, “Introduction,” in *The Arc of the Moral Universe and Other Essays* (Cambridge: Harvard University Press, 2010), p. 2.

its own “freely willed” maintenance—traceable to citizens’ sense of justice, rather than manipulation or coercion—in the face of countervailing forces in human nature. I argue against stability’s justificatory significance, contending that it defies what it is to hold a moral conviction to dilute or abandon it in the face of empirical claims about the likelihood of its realisation. I nevertheless explain that stability retains crucial normative importance for two reasons. Firstly, stability matters, plain and simple, because justice matters. To care about justice is to care about its enduring realisation, and thus to attend to the creation of arrangements that render its realisation not merely possible, but probable. The second reason to care about the freely willed achievement of justice concerns its superior moral quality to arrangements in which individual rights are protected through coercion or manipulation. Because I shall argue that justice appraises not mere distributions of freedom and wealth, but rather the character of the underlying social relationships—relationships constituted by certain enduring attitudes and their manifestation in enduring patterns of conduct—such relationships *must* be freely willed in order to be fully just.

These two reasons together motivate a replacement of Rawls’s version of the stability test—according to which stability is some justificatory hurdle set by reason on our convictions’ justification—to my version of the stability test—according to which stability becomes a practical challenge set by our own convictions *for us*. That practical challenge is to answer the following question: *how shall we fortify our capacity for justice against the forces within our nature that threaten to destabilise it, thereby empowering ourselves to identify accurately, and freely comply with, our fundamental duties to respect one another as free and equal?* I shall call citizens’ provisional answer to that question their *stability charter*, comprising various *fortificational measures* that citizens have duties to sustain. The remainder of the dissertation is devoted to the proposition that several familiar practices within liberal-

democratic political life are best conceived as examples of these fortificational measures, and that interpreting them accordingly helps us understand how they are best justified.

Chapter Three argues that practices of democratic discourse, especially those involving engagement with citizens whose specific interests are different from one's own, are partially justified by their role in increasing the reasonableness of citizens' substantive convictions. Citizens are subject to a fortificational duty to undertake democratic reasoning in a way that is organised to improve the moral reasoning of all. The chapter argues that this is a plausible way to understand the widely held conviction that democracy's value is partially epistemic, and it discusses a proposal for how democratic practices might secure their epistemic value.

Chapter Four argues that citizens are required to design institutions that militate against certain predictable threats to the morally successful operation of democratic processes. Focusing as a case study on the practice of rights-based judicial review of legislation, the chapter defends judicial review as an institution that democratic citizens and their officials have reason to create or sustain for the purpose of disciplining democracy in the face of citizens' enduring susceptibility to moral failure. Judicial review, when functioning properly, is not opposed to democracy; rather, it serves as a didactic crutch that helps to keep democracy oriented toward the achievement of justice. However, it can only serve this role—without falling prey to the most compelling scholarly objections against it—when the injustice of the legislation it condemns is beyond reasonable dispute.

Chapter Five explains how the practice of criminal punishment is fruitfully understood as one element of citizens' fortificational infrastructure. It argues that criminal offenders have special fortificational duties to undertake extra measures to fortify their sense of justice in the aftermath of its evident malfunction, and it explains

how the enforcement of these duties nevertheless respects offenders as moral agents. These “fortificational punishments” include demanding assignments enjoining introspection on why agents permitted themselves to act wrongly, and the development of strategies for improved future compliance.

Finally, Chapter Six examines a class of citizens that pose special threats to the enduring achievement of justice: “unreasonable citizens” who reject the political liberal understanding of justice. The chapter argues that the task of convincing such citizens to abandon their unreasonable convictions is one belonging to all citizens, who are subject to duties to convince those over whom they are likely to exert a productive influence to alter the content of their normative beliefs. It specifies and defends one important implication of this duty: that reasonable citizens belonging to the same religious and cultural communities ought to engage their unreasonable counterparts to explain why their shared comprehensive doctrine should not be interpreted in a way that yields illiberal commitments.

These explorations do not propose to offer an exhaustive account of what a stability charter includes. My claim is not the ambitious one that *my* views on punishment and judicial review, for example, ought to belong to an overlapping consensus, since they are as amenable to reasonable disagreement as theories of distributive justice. Rather, my aim is more modest: a citizenry’s public philosophy properly exhibits a *shared concern* with stabilising the public’s sense of justice, and a shared commitment to regard reasonable measures pursued to that end (such as those sketched in this thesis) as morally legitimate when pursued democratically. A serious concern with justice, in short, is not merely a concern with identifying substantive standards of treatment and then *hoping* that we will be strong enough to act in accordance with them. Given our fallibility, it also entails a commitment to designing and sustaining arrangements that increase the likelihood of its

enduring realisation—not ignoring our own fallibility, but reckoning with it. History shows us how easy it is for human beings to be morally subverted by pernicious beliefs and temptations. By recognising our own fallibility and taking proactive steps to fortify our fidelity to one another, we affirm, not disparage, our capacity for justice.

CHAPTER ONE

Liberal Foundations

In this chapter I introduce the normative framework of the dissertation by describing the foundational elements of a *commitment to justice* as conceived within John Rawls's contractualist theory of political liberalism. To have a commitment to justice, on this view, is, firstly, to view one's fellow citizens in accordance with a particular understanding of moral personality that regards them as *free and equal*—as moral agents of equal worth who ought to be respected as capable of framing, revising, and pursuing their chosen conception of a good life; secondly, to identify terms of interaction that are mutually acceptable among persons so understood; thirdly, to specify *democratically* the precise demands of those terms; and fourthly, to regulate one's relations with others accordingly. My limited aim here is to prepare the arguments to come in future chapters by drawing attention to particular elements of the view to which I shall frequently refer. We cannot know what it is to fortify our sense of justice until we know what our sense of justice properly directs us to pursue.

In Section 1, I present the political liberal conception of moral personality and explain why it properly regulates the terms of citizens' relationships with one another. I also explain that the purpose of these relationships is to generate cooperatively the collective goods that enable citizens to develop and exercise their moral powers. In Section 2, I make some preparatory comments about the procedure of reasoning by which citizens ought to arrive at determinate judgements about the terms that govern their relationships, and in Section 3, I make some preparatory remarks on the significance of democracy for my account. I conclude by remarking on the content of citizens' ideal common ground in a well-ordered liberal democracy.

§1. The Ideal of Justice

1.1. The political liberal conception of moral personality

The contractualist framework of political liberalism is a doctrine about what we owe to each other, morally, when setting the basic terms of our relations as citizens. Its distinctive appeal, I believe, lies not in the familiarly celebrated basic liberties it affirms—many theories, after all, endorse individual rights—but rather in how it arrives at them. According to John Rawls, the moral imperative to respect people in the ways familiarly enjoined by self-professed liberals is not to be explained by its contribution to aggregate utility, its prudential rationality for self-interested actors, or its conformity with a single conception of the good that all persons should be expected to share.¹ Rather, the distinctive appeal of the familiar “basic liberties” is that they flow from a certain widely embraced understanding of how we ought to regard and relate to one another: an understanding of one another as *free and equal*.² This is not an alien understanding to

¹ For utilitarian approaches to liberalism, see Jeremy Bentham, “Of the Principle of Utility,” in *Bentham’s Political Thought*, edited by Bikhu Parekh (London: Croom Helm Limited, 1973), and Robert E. Goodin, *Utilitarianism as a Public Philosophy* (Cambridge: Cambridge University Press, 1995). For the seed of a prudential defence of liberalism, see Thomas Hobbes, *Leviathan* (London: Penguin Books, 1968), and for a contemporary update, see David Gauthier, *Morals by Agreement* (Oxford: Oxford University Press, 1986). For the idea that liberalism is justified on the basis of a particular conception of a good life that ought to be led by all—for example, a conception of the good according to which human excellence is achieved through experimentation, rejection of tradition, and autonomous self-development—see J. S. Mill, *On Liberty*, edited by Elizabeth Rapaport (Indianapolis: Hackett Publishing Company, 1978). For a more plausible variant of that view, according to which liberalism is justified because it enables the pursuit of various worthwhile conceptions of the good, see Joseph Raz, *The Morality of Freedom* (New York: Oxford University Press, 1986). For another influential expression of the view that liberalism is best understood as following from, or otherwise necessarily connected to, true beliefs about the good, see Ronald Dworkin, “Foundations of Liberal Equality,” in *The Tanner Lectures on Human Values*, vol. 11, edited by G. B. Peterson (Salt Lake City: University of Utah Press, 1990), pp. 2f, and *Justice for Hedgehogs* (Cambridge: Harvard University Press, 2011).

² The most important developments of the characterisation of persons as free and equal in modern political philosophy are found in John Locke, *The Second Treatise of Government*, edited by C. B. Macpherson (Indianapolis: Hackett Publishing Company, 1980), and Jean-Jacques Rousseau, “On the Social Contract,” in *The Basic Political Writings* (Indianapolis: Hackett Publishing Company, 1987). The idea that regarding persons as free and equal in some way involves acting in ways that are acceptable to them in virtue of that status is developed most

members of liberal democracies, but one which we already affirm, rooted in the twin slogans that each person has his own life to lead, and that each person is equally valuable. Rawls contends that the aspiration to see others in this way is a deeply embedded feature of our existing political and moral practice, one to which we are reflectively committed; this is the significance of his frequently misunderstood contention that the idea of persons as free and equal is latent in the “public political culture of a democratic society,” “in its basic political texts (constitutions and declarations of human rights),” and in our “common practical reason.”³ This vision of personhood is deeply woven into our moral phenomenology. It is presupposed by the substantive attitudes and practices that characterise quotidian moral criticism—of public officials and fellow citizens—in free societies. The “reactive attitudes”⁴—both self-regarding (pride, guilt, shame) and other-regarding (praise, blame, resentment)—that comprise the substance of ordinary political and moral life all presuppose a certain understanding of the status of all that makes such attitudes appropriate. In liberal democracies, that status is a certain view of persons as entitled to be treated as equally valuable and entitled to chart their own course in life, consistent with the rights of others to do the same.

On Rawls’s particular interpretation of this widely affirmed view, the normative self-understanding that lies at the heart of our civic phenomenology—the view of persons as free and equal—is one that regards all as in possession of *two moral powers*:

significantly in Immanuel Kant, “Groundwork of the Metaphysics of Morals,” in *Practical Philosophy*, edited by Mary Gregor (New York: Cambridge University Press, 1996).

³ John Rawls, *Justice as Fairness: A Restatement* (Cambridge: Harvard University Press, 2001), 19; John Rawls, *Political Liberalism*, p. 90. For misreadings of this idea, see, for example, Richard Rorty, “The Priority of Democracy to Philosophy,” in *Objectivity, Relativism, and Truth* (Cambridge: Cambridge University Press, 1991), p. 189. Here I agree with the criticism of Rorty advanced in Jürgen Habermas, *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*, translated by William Rehg (Cambridge: The MIT Press, 1998), p. 62.

⁴ P. F. Strawson, “Freedom and Resentment,” in *Freedom and Resentment and Other Essays* (Abingdon: Routledge, 2008), p. 6.

firstly, a capacity to frame, revise, and pursue a conception of the good; and secondly, a capacity to see the force of moral reasons and act on them—what Rawls calls *the sense of justice*. These powers of practical reasoning—which Rawls terms *rationality* and *reasonableness*—are ones we implicitly regard others as possessing when we relate to them in political and moral life. “Reasonable and rational agents are normally the units of responsibility in political and social life and may be charged with violations of reasonable principles and standards.”⁵ Insofar as others meet the threshold for possessing these powers—so long as they possess this “range property”⁶—we regard them as equally valuable, entitled to be affirmed as free, and subject to moral expectations to regard others accordingly. (I shall sometimes use the term *free equals* as shorthand.) The central insight of political liberalism, according to my preferred interpretation, is that a proposed social norm or rule is *publicly justified*—properly regarded as morally binding on all—only when it can be justified in accordance with an understanding of persons as free equals.⁷

⁵ Rawls, *Political Liberalism*, p. 50.

⁶ Rawls, *A Theory of Justice*, p. 444ff.

⁷ The other major developer of political liberalism is Charles Larmore. See especially his *Patterns of Moral Complexity* (Cambridge: Cambridge University Press, 1987), “Political Liberalism,” *Political Theory* 18, 3 (1990): 339-360, *The Morals of Modernity* (Cambridge: Cambridge University Press, 1996), “and “The Moral Basis of Political Liberalism,” *The Journal of Philosophy* 96, 12 (1999): 599-625. Other exponents include Bruce Ackerman, *Social Justice and the Liberal State* (New Haven: Yale University Press, 1980); Martha Nussbaum, *Women and Human Development* (Cambridge: Cambridge University Press, 2000), *Frontiers of Justice* (Cambridge: Harvard University Press, 2006), and “Political Liberalism and Perfectionist Liberalism,” *Philosophy & Public Affairs* 39, 1 (2011): 3-45; Thomas Nagel, “Moral Conflict and Political Legitimacy,” *Philosophy & Public Affairs* 16 (1987): 215-40, and *Equality and Partiality* (New York: Oxford University Press, 1991); and Jonathan Quong, *Liberalism Without Perfection* (New York: Oxford University Press, 2011). Jeremy Waldron invites an understanding of himself as a political liberal in “Theoretical Foundations of Liberalism,” in *Liberal Rights: Collected Papers 1981-1991* (Cambridge: Cambridge University Press, 1993), especially at pp. 36-7, but he has since resisted this classification.

Consider also three further examples that do not fit the Rawlsian understanding of political liberalism but contain vital elements of it. Firstly, Gerald Gaus is sometimes regarded as an exponent of this view, though his relativist understanding of what is involved in treating others as free and equal—according to which public justification is always indexed to idealised versions of existing persons’ belief sets—is so markedly different from Rawls’s that it is properly classified as a distinctive approach in its own right. See his *Justificatory Liberalism: An Essay on Epistemology and Political Theory* (New York: Oxford University Press, 1996), *The Order of Public Reasons: A Theory of Freedom and Morality for a Diverse and Bounded World* (New York: Cambridge

The commitment to regarding persons as free equals is plainly “non-neutral” to the degree that it contradicts rival understandings of persons that see persons, for example, as meriting differential treatment by virtue of skin colour or social caste. While it rules out such illiberal understandings of persons, however, it is perfectly compatible with any deeper account of the nature of personhood and the deepest source of moral value.⁸ Different “comprehensive” philosophical, ethical, and religious doctrines can come to affirm the political liberal conception of moral personality in their own way, joining together to form an “overlapping consensus” upon it and its implications for our social life.⁹ Note that the compatibility of liberal ideals with religious and philosophical diversity is not to be understood as some fortuitous convenience. It flows from what it *is* to see and regard others as free equals: as the kinds of persons entitled to draw their own conclusions about the vexed and difficult questions about the meaning and nature of life.¹⁰ That such questions *are* vexed and difficult is manifest to anyone who thinks

University Press, 2011). Secondly, Jürgen Habermas has developed a sophisticated philosophical architectonic centred around the ideal of free communication oriented to securing mutual moral understanding on matters of justice. While Rawls himself admits close similarities between their views, he nevertheless regards Habermas’s theory as a “comprehensive doctrine” due to its controversial metaethical and metaphysical commitments; Rawls, “Reply to Habermas,” p. 376. (Given its lack of a view on the good, a more accurate charge would have been “partially comprehensive doctrine.”) See Habermas, *Moral Consciousness and Communicative Action*, translated by Christian Lenhardt and Shierry Weber Nicholsen (Cambridge: The MIT Press, 1999), *Between Facts and Norms*, and *The Inclusion of the Other*, edited by Ciaran Cronin and Pablo De Greiff (Cambridge: The MIT Press, 1998). Finally, T. M. Scanlon’s influential contractualist view of ethics has deep affinities with Rawls’s views, especially his view of motivation, as Rawls himself says; see *Political Liberalism*, p. 49n. However, Scanlon aspires to understand a wider subject matter than justice, including such issues as the wrongness of breaking promises. See Scanlon, “Contractualism and Utilitarianism,” in Amartya Sen and Bernard Williams (eds.), *Utilitarianism and Beyond* (Cambridge: Cambridge University Press, 1982), and *What We Owe to Each Other* (Cambridge: Harvard University Press, 1998).

⁸ This renders mistaken the central argument advanced in Michael Sandel, *Liberalism and the Limits of Justice*, second edition (Cambridge: Cambridge University Press, 1998).

⁹ Rawls, *Political Liberalism*, pp. 144ff. The idea of the overlapping consensus arises in Chapters Two and Six of this dissertation.

¹⁰ Thus an important criticism that political liberalism levies at perfectionist liberalism is that the latter is *itself* illiberal: by insisting that we structure our shared affairs around the quest to improve people’s compliance with a correct or meaningful conception of a good life, a political order so conceived disrespects their capacity to reason freely and choose for themselves, engaging in

seriously about the sources of pluralism: those of comparable intellectual might who reason with thoughtfulness and integrity about fundamental questions of nature and value are bound to disagree, given what Rawls calls “the burdens of judgement.”¹¹

Liberal politics, on Rawls’s view, is the politics of people who feel the grip of this self-understanding and are committed to regulating their civic relationships by it—who, despite their divergent underlying views about the meaning and nature of life and the universe, find themselves moved to believe that the understanding of persons as free equals ought to be our “lodestar” in determining how we ought to structure and regulate our relationships with one another as fellow citizens.¹² The ideal of the liberal society is the collective organisation of such *reasonable citizens*. I shall say more in Chapter Six about those who dispute liberal ideals—those, that is, who find the understanding of persons as free equals deeply alien to their moral sensibility. But it is important to note that those who find themselves moved by liberal ideals do not regard their commitment as merely preferential in character. As the reactive attitudes of resentment and blame make clear, when others do not comport their behaviour with certain standards we regard as binding, our reaction is not one of simple disappointment or indifference—like finding out that our best friend happens to disagree with us about the best flavour of ice cream. For those who believe in those standards, the lack of compliance with them by others is best understood—and is unshakeably perceived to be—a *moral failure*. And the political liberal conception of moral personality helps us see why this reaction is appropriate: because

“expressive subordination.” See Nussbaum, “Political Liberalism and Perfectionist Liberalism,” p. 35; and Quong, *Liberalism Without Perfection*, especially pp. 73-107.

¹¹ The burdens of judgement concern the difficulties in reasoning that lead to disagreements even among those who reason with great conscientiousness. See Rawls, *Political Liberalism*, pp. 56-57.

¹² Stephen Macedo, *Liberal Virtues: Citizenship, Virtue, and Community in Liberal Constitutionalism* (Oxford: Clarendon Press, 1990), p. 78.

such behaviour reflects a failure to value persons in the right way. Relating to others as free and equal is a moral requirement binding on all, not a worthy but optional ideal.

1.2. Ideal social cooperation as the realisation of just social relationships

I have presented the basic contours of the political liberal conception of moral personality. For liberals, this conception furnishes the practical orientation within which one confronts the practical question of how we ought to relate to our fellow citizens. So stated, however, the view remains indeterminate. Surely we should relate to *all* human beings in the manner presupposed by the political liberal conception of moral personality; surely those persons with whom I share a polity by accident of birth are not in virtue of that contingency worthy of being regarded as equally valuable moral agents with their own lives to lead, while others are not? This emphasis on *citizens'* relations reflects the distinctive *purposes* of one's relations with fellow citizens that do not characterise one's relations with non-citizens.¹³ But what is the right conception of relations between citizens?

The answer I shall assume for the purposes of this dissertation is this: the purpose of persons' relationships with one another *qua* fellow citizens is to generate cooperatively the collective goods that enable the development and exercise of citizens' two moral powers. On this view, the purpose of our "civic relationship" is to work together to produce and sustain a social world, that, firstly, enables citizens to see themselves as moral agents moved by the value of others, and that, secondly, secures us with "a central class of collective goods, namely those goods necessary for developing

¹³ "A moral conception incorporates a conception of the person and of the relations between persons." Rawls, "The Independence of Moral Theory," in *Collected Papers*, edited by Samuel Freeman (Cambridge: Harvard University Press, 1999), p. 293.

and acting on a plan of life.”¹⁴ These “primary goods”—goods such as “basic rights and liberties,” “freedom of movement and free choice of occupation against a background of diverse opportunities,” and “income and wealth,” among others—are indispensable for beings with the capacity for framing, revising, and pursuing a particular conception of good.¹⁵ We require this social world to realise our capacities because of our nature as embodied, social beings.¹⁶ The terms of cooperation that successfully realise such a world are *the principles of justice*.¹⁷ If persons cooperate in this way—if they regulate their relation by the relevant terms—“what their cooperation achieves (in the ideal case) is a just basic structure... Their cooperation is to assure one another political justice.”¹⁸

This interpretation of the point of citizens’ relationships—to maintain a society with the requisite *basic structure* (“the political constitution and the principal economic and social arrangements”¹⁹) to enable the exercise of citizens’ defining capacities—is philosophically controversial. I shall not defend it here, though I believe it is the right

¹⁴ Andrea Sangiovanni, “Global Justice, Reciprocity, and the State,” *Philosophy & Public Affairs* 35, 1 (2007): 3-39, p. 4. Sangiovanni’s argument at times implies that the pursuit of a just basic structure is a morally optional pursuit that, once undertaken, grounds people’s claims for a “fair return” on the work they have put into it; p. 27. This invites us to see the theory of social justice as an extension of the theory of fair play, which I seek to resist. Sangiovanni notes this issue, but appears agnostic on it.

¹⁵ Rawls, *Political Liberalism*, p. 181.

¹⁶ I shall not elaborate this point here, but I mean it in the most uncontroversial possible sense: people cannot exercise their capacities as free and equal without the right kind of supporting context. “[G]iven our dependence on society, we could not be this sort of person unless institutions developed and encouraged our capacity so to act and others publicly to acknowledge its realization. *People’s attaining this conception of the person would be the achievement of social cooperation; for success depends on social forms and mutual recognition.*” Rawls, “The Independence of Moral Theory,” pp. 293-94. Emphasis added.

¹⁷ “Think, then, of the principles of justice as designed to form the social world in which our character and our conception of ourselves as persons, as well as our comprehensive views and their conceptions of the good, are first acquired, and in which our moral powers must be realized. These principles must give priority to those basic freedoms and opportunities’ in background institutions of civil society that enable us to become free and equal citizens in the first place, and to understand our role as persons with that status.” Rawls, *Political Liberalism*, p. 41.

¹⁸ *Ibid.*, p. 42n.

¹⁹ Rawls, *A Theory of Justice*, p. 6

interpretation of Rawls; it rests in concord with both Rawls's specification of a natural duty to support just institutions, and his controversial "duty of assistance," according to which human beings everywhere owe it to one another to ensure that everyone is part of *some* social association that nurtures their basic capacities and enables their exercise, even if not to the full extent as in a fundamentally just society.²⁰ It also immunises Rawls's view from the morally unappealing interpretation of his work that reads him as simply engaging in a moralised anthropological reconstruction of existing social practices.²¹ This interpretation is "cosmopolitan" insofar as it affirms the equal moral value of each human being, but I leave open the question of whether it properly belongs to our long-term moral ideal that the world be governed by a single basic structure—a single set of terms of interaction that regulate the relationships of all human beings as co-citizens. For the purposes of this dissertation, however, I shall focus on the relationships between citizens of a single social order.

Three further comments are necessary. Firstly, our duties to our fellow citizens are distinctive because we owe it to them to participate in the maintenance of a particular social order where people are supported in the realisation of their moral capacities. But it must not be ignored that we *also* owe our fellow citizens the same "natural duties" that we owe everyone in virtue of their status as human beings. These are duties "which hold between persons irrespective of their institutional relationships; they obtain between all as equal moral persons."²² These include familiar negative duties against harming others,

²⁰ This. See Rawls, *The Law of Peoples*, pp. 83ff.

²¹ Aaron James, "Constructing Justice for Existing Practice: Rawls and the Status Quo," *Philosophy & Public Affairs* 33, 3 (2005): 281-316. The problem with James's view is that he supposes we determine the purposes of social cooperation by examining what existing societies do and who their participants are. It therefore forfeits the *normativity* of Rawls's thesis: we *must* create and sustain just basic structures.

²² "The principles that hold for individuals, just as the principles for institutions, are those that would be acknowledged in the original position." Rawls, *A Theory of Justice*, p. 99.

as well as positive duties of assistance. Some are keen to restrict the term “justice” to the former class of duties that bind us as co-citizens, and refer to the latter duties as “humanitarian,” but I will not adopt this terminological distinction here.²³ There is a venerable tradition in liberal political philosophy of thinking that Fred acts *unjustly* toward Roger when he kills him, and I shall follow this tradition. I do so without wishing to deny that the phrase “just society” refers to a special sort of achievement (the achievement of, we might say, *social justice*). We can simply say: *The commitment to justice is a commitment to viewing all as free and equal in our relations with them, which has special implications for our relations with our fellow citizens.* This seems to me to be exactly the right way to express the universalistic sentiment of liberalism in a way that accounts for Rawls’s insistence on the importance of concrete “social forms” for the development and exercise of the moral powers.²⁴

Secondly, recall that we began by asking: given that we view one another as free and equal, how shall we conduct our relationships as fellow citizens? It is often said that the subject matter of (social) justice is institutions. Yet it is important to see that institutions do not refer some collection of skyscrapers and offices, but to *public systems of rules*, the existence of which is marked by “the realization in the *thought and conduct* of certain persons at a certain time of the actions specified by those rules.”²⁵ In other words, just social institutions simply *are* the upshot of citizens’ choices²⁶ to regulate their

²³ Thomas Nagel, “The Problem of Global Justice,” p. 119.

²⁴ “People’s attaining this conception of the person would be the achievement of social cooperation; for success depends on social forms and mutual recognition.” Rawls, “The Independence of Moral Theory,” pp. 293-94.

²⁵ Rawls, *A Theory of Justice*, p. 48. Emphasis added.

²⁶ This claim does not take sides in the debate about *which* choices are thereby properly regarded as regulated under the ambit of social justice. It is therefore agnostic about the debates pursued in Cohen, *Rescuing Justice and Equality*, and the related questions pursued in Liam Murphy, “Institutions and the Demands of Justice,” *Philosophy & Public Affairs* 27, 4 (1999): 251-91. But no one disputes that people’s choices to pay their *taxes* are enjoined by distributive justice.

relationships with one another on the required terms.²⁷ I draw out this point for two reasons. Firstly, we must always remember that justice can never be achieved and then left on auto-pilot to take care of itself. There is no “perfect technology of justice,”²⁸ only the choices of “living, breathing persons”²⁹ who can decide at any time whether to continue coordinating their actions in the manner that produces a just social order, or to give it up. Secondly, by insisting that justice enjoins us to undertake our *relationships* with one another in a certain way—to produce a just social order—we are drawing attention to an indispensable element of the political liberal view at its most attractive, owing to what a relationship *is*. Relationships are constituted, firstly, by certain reciprocal attitudes, and secondly, by manifestations of those attitudes in patterns of appropriately motivated conduct. For that reason, justice is not agnostic to the attitudinal dimensions of political and moral life. It does not merely say: “provide everyone with the quantum of non-interference and material resources that they are owed.” It enjoins us to *respect* them as free and equal; claims about just distributions flow *from* that ideal.³⁰ When we ask what distinguishes a social world in which people live together as free and equal from one in which they live together as masters as slaves, our analytical gaze fastens *not* on any discrepancy in the “pattern in the distribution of divisible goods,” but on the character of

²⁷ For defences of this view, see Elizabeth Anderson, “What Is the Point of Equality?” *Ethics* 109, 2 (1999): 287-337; Joshua Cohen, “Democratic Equality” *Ethics* 99, 4 (1898): 727-51; Samuel Scheffler, “What Is Egalitarianism?” *Philosophy & Public Affairs* 31, 1 (2003): 5-39; Jonathan Wolff, “Fairness, Respect, and the Egalitarian Ethos,” *Philosophy & Public Affairs* 27, 2 (1998): 97-122. Note that the idea of relational egalitarianism is different from the sense of “relational” used in Derek Parfit, “Equality and Priority,” *Ratio* 10, 3 (1997): 202-21, where Parfit deems “relational” those views that hold that it is intrinsically valuable for people to possess the same quantum of some good.

²⁸ Bruce Ackerman, *Social Justice and the Liberal State*, p. 21.

²⁹ Eric Beerbohm, *In Our Name: The Ethics of Democracy* (Princeton: Princeton University Press, 2012), p. 2.

³⁰ See Christian Schemmel, “Why Relational Egalitarians Should Care About Distributions,” *Social Theory and Practice* 37, 3 (2011): 365-90.

their relationships.³¹ Only when citizens *affirm* each other in this way, rather than grudgingly comply with justice's directives, "can there be true community between persons in their common practices."³² The relational explanation accounts for our considered convictions that justice is not indifferent toward our attitudes.

Finally, I want you to notice that the exploration thus far has made *no* mention whatsoever of the idea of *coercion*. It has simply not been necessary to refer to coercion to illuminate any of political liberalism's core ideas. This may seem puzzling. Coercion is a central feature of actual political life: existing states threaten punishment for those who contravene its directives. Rawls frequently refers in his later work to state power as *coercive* power.³³ Charles Larmore goes so far as to say that the justification of coercion is the defining topic of political philosophy.³⁴ Recent work on the theory of global justice suggests that the defining "function of justice...is to evaluate the moral legitimacy of state coercion."³⁵ Nevertheless, without denying that coercion is presumptively wrongful and thus ought not to be undertaken unless morally justified, I believe that fastening our philosophical gaze tightly upon coercion seriously distorts our thinking about justice for two reasons.

³¹ Anderson, "What's the Point of Equality," p. 336. It is sensible to understand justice as divisible between *nonrelational* goods and *relational* goods, and still think of the question distributively, but I do not pursue this approach here. See Zofia Stemplowska, "Responsibility and Respect: Reconciling Two Egalitarian Visions," in *Responsibility and Distributive Justice*, edited by Carl Knight and Zofia Stemplowska (New York: Oxford University Press, 2011), p. 116.

³² Rawls, "Justice as Fairness," in *Collected Papers*, p. 59. Scanlon flags this passage in his "Rawls on Justification," in *The Cambridge Companion to Rawls*, edited by Samuel Freeman (New York: Cambridge University Press, 2003), p. 161. I am admittedly using this passage for slightly different purposes; Rawls is emphasizing the importance that citizens share the same basic convictions about justice. But his formulations resonate with the relational interpretation.

³³ Rawls, *Political Liberalism*, pp. 139-40.

³⁴ "Justice as such cannot be the state's concern. It has to be justice insofar as it can be part of an authoritative set of rules that will be [coercively] binding on all." Charles Larmore, "What Is Political Philosophy?" *The Journal of Moral Philosophy* 10 (forthcoming 2013): 1-33, p. 10. See also Larmore, "The Moral Basis of Political Liberalism," p. 607.

³⁵ Laura Valentini, *Justice in a Globalized World* (New York: Oxford University Press, 2011), p. 124.

Firstly, a focus on coercion invites us to circumscribe justice's ambit in implausible ways. Respect for others as free and equal demands many things of us, and it is not clear they are all coercively enforceable. For example, the systematic decisions of parents to spend all their time helping their sons with their homework, instead of their daughters, strikes us as deeply unjust—despite the fact that, all-things-considered, reasons of justice count against sending police officers to monitor parents' conduct.³⁶ More strikingly, pervasive racist attitudes, even if the strictures of non-discrimination law are complied with tightly, deeply undermine the moral quality of a social order.³⁷ Consider, too, that the refusal of white supremacists to fulfil “the duty of civility”—the duty to offer reasons in public discourse that one sincerely believes are acceptable to all as free and equal—strikes us as morally wrong, even if they never stand a chance of influencing coercive law.³⁸

Secondly, casting official coercive rules as the defining site of justice encourages a perverse normative orientation toward the phenomenon of state coercion—a result that exposes coercion's limited normative significance. An increasingly popular class of views in the literature on global justice supposes that the reason citizens are owed justice by their state—for example, in the form of redistributive transfer payments—is because, *inter alia*,³⁹ they are *coerced* by their state. Because coercion is “presumptively wrong,”⁴⁰ according to Michael Blake, it triggers a heightened form of scrutiny. Using the example

³⁶ For a general discussion of injustice in the family, see Susan Moller Okin, *Justice, Gender, and the Family* (New York: Basic Books, 1989). For reflection on the value of the family within liberal political philosophy, see Harry Brighouse and Adam Swift, “Legitimate Parental Partiality,” *Philosophy & Public Affairs* 37, 1 (2009): 43–80.

³⁷ See Miranda Fricker, *Epistemic Injustice: Power and the Ethics of Knowing* (New York: Oxford University Press, 2007).

³⁸ On the duty of civility, see Rawls, *Political Liberalism*, p. 444.

³⁹ On the complex view of Thomas Nagel, coercion must be supplemented by the provision that the state is speaking “in the name” of those whom it coerces. Nagel, “The Problem of Global Justice,” p. 129.

⁴⁰ Michael Blake, “Distributive Justice, State Coercion, and Autonomy,” *Philosophy & Public Affairs* 30 (2001): 257–96, p. 278.

of coerced taxation, Blake argues that so long as coercion “could not reasonably be rejected by those who face the taxation,” it is acceptable. Blake’s reasoning, however, gets the order of justification exactly backwards. We conceive of something as morally worthwhile or morally required, and *then* we proceed to ask whether it ought to be enforced coercively. Coercion does not *trigger* our fundamental duties of justice in the way Blake’s description implies.⁴¹ Rather, it is simply a mechanism human beings employ to enforce them.⁴²

Andrea Sangiovanni advances a similar argument against coercion-focused theorists with the following example, which it is worth quoting in full:

Let us...suppose that all local means of law enforcement—police, army, and any potential replacements—are temporarily disarmed and disabled by a terrorist attack. Suppose further that this condition continues for several years. Crime rates increase, compliance with the laws decreases, but society does not dissolve at a stroke into a war of all against all. Citizens generally feel a sense of solidarity in the wake of the attack, and a desire to maintain public order and decency despite the private advantages they could gain through disobedience and noncompliance; this sense of solidarity is common knowledge and sufficient to provide assurance that people will (generally) continue to comply with the law. The laws still earn most people’s respect: the state continues to provide the services it always has; the legislature meets regularly; laws are debated and passed; contracts and wills drawn up; property transferred in accordance with law; disputes settled through legal arbitration, and so on. Does the legal system no longer require the same kind of justification to those most disadvantaged by it simply because it now lacks coercive power?⁴³

The answer Sangiovanni encourages is “no,” and I think he is right about that. I draw attention to Sangiovanni’s example for a fresh purpose, however. Once we see that

⁴¹ “Consider the structurally analogous case of interpersonal interference: if you interfere with my choice by issuing a threat to the effect that you will shoot me in the leg if I carry through with my intended action, we would not say that I have an *ex post* claim on your interference that it only be employed for moral purposes; we would ask what your *ex ante reason* for interference was – and here it makes a world of difference whether you interfered with my attempt to rape and kill some or my attempt to visit my mother.” Malte Frøslee Ibsen, “Justice as Non-Domination” (unpublished manuscript). I thank Malte Ibsen for this example and for many discussions on this issue.

⁴² See related arguments in Andrea Sangiovanni, “The Irrelevance of Coercion, Imposition, and Framing to Distributive Justice,” *Philosophy & Public Affairs* 40, 2 (2012): 79-110.

⁴³ Sangiovanni, “Global Justice, Reciprocity, and the State,” pp. 10-11.

coercion could return to the society Sangiovanni describes only due to the conscientious choices of its members—who establish and continually choose to sustain coercive practices—we see coercion for what it really is: simply one mechanism among many that we may or may not decide to use to enforce our convictions about justice. That is not how it sometimes *feels* when we are coerced by a state. But that is how the structure of the justification should be understood. Whether we need coercion, and why we might, will concern me in due course.⁴⁴

§2. Contractualism and Public Reason

Justice demands that we conduct our relationships with our fellow citizens in accordance with the political liberal conception of moral personality, working together to provide the goods that enable all to pursue their respective conceptions of the good. The political liberal conception of moral personality thereby issues particular terms of cooperation by

⁴⁴ This thesis may appear in conflict with two widely affirmed contentions. The first is that a state is *definitionally* a coercive entity; the second is that *law* is definitionally a coercive phenomenon. The account here resists both characterisations. Max Weber's proposition that states are entities that each possess a "*monopoly of the legitimate use of physical force* within a given territory" has been extremely influential in political science. See Max Weber, "Politics as a Vocation," in *From Max Weber: Essays in Sociology*, edited and translated by H. H. Gerth and C. Wright Mills (New York: Oxford University Press, 1946), p. 78. Emphasis is Weber's. As Leslie Green notes, "the view that the state is essentially a coercive order is unsupported by the evidence." Leslie Green, *The Authority of the State* (New York: Oxford University Press, 1998), p. 71. But even if it were supported by the evidence as an *empirical* characterisation of how citizens regard their states, the normative understanding of a state belonging the Rawlsian basic structure—as part of the social structure that citizens cooperatively sustain to enable the exercise of their moral powers—does not *itself* make reference to the idea of coercion. Now it may appear that insofar as states make *laws*, and *law* by its nature consists in coercive commands, then the idea that the state is essentially coercive remains. But as H. L. A. Hart has persuasively argued against both John Austin and the sophisticated variation defended by Hans Kelsen, the understanding of law as essentially coercive is also vulnerable to powerful objections.⁴⁴ I cannot engage in these jurisprudential disputes here. I invoke them simply to clarify that my position is not remotely idiosyncratic outside of contemporary political theory. H. L. A. Hart, *The Concept of Law*, second edition (New York: Oxford University Press, 1994), pp. 18-78; See also John Austin, *The Province of Jurisprudence Determined and the Uses of the Study of Jurisprudence* (Indianapolis: Hackett Publishing Company, 1998); and Hans Kelsen, *General Theory of Law and State*, translated by Anders Wedberg (Cambridge: Harvard University Press, 1945; reprinted in Clark: The Lawbook Exchange Ltd., 2007).

which we must regulate our conduct. But how do we arrive at conclusions about what those terms actually are?

We can arrive at the answer to this question by asking another: in what way is political liberalism a “contractualist” doctrine? The answer to this question helps us see how we reason from the liberal conception of moral personality to reach specific conclusions about what we morally ought to do. While Scanlon seemingly aspires to reserve the term “contractualism” for his own distinctive ethical theory,⁴⁵ it has been embraced by a wide variety of political and moral philosophers to describe any theory according to which the correct principles of justice or morality can be conceived as the conclusions to some kind of agreement. That idea is vague, of course, and different contractualist theories cash it out in different ways.

The basic idea is old, but its most influential initial manifestation in early modern political thought comes in the writings of Thomas Hobbes. By imagining how prudentially rational agents in a pre-political state of nature—equally concerned to avoid the prospect of violent death—would join in agreement together to create a powerful sovereign to protect the fundamental interests of all, Hobbes teaches us how it is prudentially rational for each of us currently living under such a sovereign to endorse its existence and power.⁴⁶ *Leviathan* placed the ideal of justification to each individual—the idea that a justified political structure is one that each person has reason to accept—on the agenda of modern political philosophy. Yet the parties to Hobbes’s hypothetical contract differ from the parties imagined by many subsequent thinkers in stipulating that their reasons for endorsing the agreed-upon principles are *prudential*. According to the versions of contractualism embraced by Locke, Rousseau, and Kant, we ought to

⁴⁵ Scanlon, *What We Owe to Each Other*, p. 5.

⁴⁶ Hobbes, *Leviathan*, p. 223ff.

conceive the relevant agreements as underwritten by parties' *moral reasons*. *Real* contractualists (as opposed to Hobbesian "contractarians," as some make the terminological cut) conceive of the principles of morality as the conclusions of agreements in which parties are, in some manner, *morally motivated*.

Political liberalism falls into this non-Hobbesian camp. On this view, the way to move from the ideal of respecting persons as free and equal to the specification of determinate moral principles is to envision an agreement between agents committed (or modelled as committed) to one another's moral status—committed, that is, to seeing one another in accordance with the liberal conception of moral personality—and inquire into the terms of interaction that such idealised contractors would consent to. According to such contractualist views, such rules derive their justificatory force from the mere fact that they would be agreed upon by appropriately described agents. "The leading idea is to establish a suitable connection between a particular conception of the person," Rawls writes, "and first principles of justice, by means of a procedure of construction."⁴⁷ Various contractualist procedures serve this role.

Rawls's *A Theory of Justice* presents the classic device of representation he calls "the original position" as a way of taking the abstract commitment to persons as free equals and focusing it into determinate principles of justice. The original position is an attempt to model formally, or "represent," the status of persons as free and equal such that determinate principles can be straightforwardly deduced. This position is an imagined agreement between agents stripped of the knowledge of their identity—as well as any moral motivation—who nevertheless seek to maximise their holdings of "primary goods," despite uncertainty as to what their actual identity is. Reluctant to agree to principles that would sacrifice his or her fundamental interests in the eventuality that he

⁴⁷ Rawls, "Kantian Constructivism in Moral Theory," in *Collected Papers*, p. 304.

or she turned out to be poorly endowed with natural talents or in possession of a minority religious or ethical view once the “veil of ignorance” is lifted, each party thereby agrees to principles that (Rawls contends) secure equal freedom, opportunity, and economic reciprocity for all. Specifically, Rawls contends that parties would agree to a set of lexically ordered principles guaranteeing, firstly, a set of equal basic liberties—such as freedom of speech, conscience, association, political participation, and bodily integrity; secondly, equality of opportunity; and thirdly, reciprocity in citizens’ economic relations, such that deviations from equal holdings of certain primary goods (such as income) are acceptable only if they work to the advantage of the worst off. These are the principles that Rawls identifies as demanded by respect for all as free equals.⁴⁸

As a heuristic device, the original position is both compelling and arresting, immediately helping citizens to secure an epistemic grip on the sorts of considerations (knowledge of race, sex, religion) that can undermine one’s attempt to arrive at mutually acceptable (or, more pithily, *fair*) conclusions. It is an easy and effective way to test a candidate principle’s universalisability; after all, the original position is “a procedural interpretation of Kant’s conception of autonomy and the categorical imperative.”⁴⁹ By focusing our reasoning in such a way that our judgements about what we ought to do must meet a test of universalisation, we ensure that the reasons that we ultimately certify as valid are reasons for *everybody*. As Christine Korsgaard writes of morality—and here we can happily substitute “justice”:

The primal scene of morality...is not one in which I do something to you or you do something to me, but one in which we do something together. The subject matter of morality is not what we should bring about, but how we should relate to one another. If only Rawls has succeeded in escaping utilitarianism, it is because only Rawls has fully grasped this point. His primal scene, the original

⁴⁸ Rawls, *A Theory of Justice*, p. 13.

⁴⁹ Rawls, *A Theory of Justice*, p. 226.

position, is one in which a group of people must make a decision together. Their task is to find the reasons they can share.⁵⁰

The reasoning of the original position is thus a form of *public reason*: an approach to the justification and deliberation of moral principles premised on the ideal that norms are justified only when they are acceptable to all—the very approach pioneered, in a very different manner, by Hobbes.

However, the indirectness of the original position has made it vulnerable to intense criticism, generating what are in my view highly distracting debates about whether the ignorance Rawls seeks is genuinely possible among “real people,” for example, or about how risk-averse parties actually are.⁵¹ Given that the methodological use of the original position is justified by the aim of providing a form of “moral geometry”⁵²—a goal Rawls concedes he does not accomplish—we are bound to ask: why go through the gymnastics of the original position when we can simply ask the fundamental question it endeavours to answer directly: *what principles respect others as free equals?* Ronald Dworkin long ago contended that the original position reflects an ideal of persons as equals, and that this idea—not any hypothetical agreement—is what matters.⁵³ And while some thinkers regard the original position as indispensable,⁵⁴ Brian Barry has

⁵⁰ Christine Korsgaard, “The Reasons We Can Share,” *Social Philosophy & Policy*, 10, 1 (1993): 24–51, pp. 24–25.

⁵¹ For the criticism that Rawls’s original position reflects substantive theses about the character of real people, see Michael Sandel, *Liberalism and the Limits of Justice*, pp. 154ff. For the claim that the Rawlsian project problematically relies on assumptions about risk (problematic in that such assumptions prevent Rawls’s theory from arriving at its desired results), see John Harsanyi, “Can the Maximin Principle Serve as the Basis for Morality? A Critique of John Rawls’s Theory,” *American Political Science Review* 69, 2 (1975): 594–606.

⁵² Rawls, *A Theory of Justice*, p. 105.

⁵³ “The original position is well designed to enforce the abstract right to equal concern and respect, which must be understood to be the fundamental concept of Rawls’s deep theory.” Dworkin, “Justice and Rights,” in *Taking Rights Seriously* (Cambridge: Harvard University Press, 1978), p. 181.

⁵⁴ For example, Paul Weithman, *Why Political Liberalism?* (New York: Oxford University Press, 2010), p. 223ff.

led the charge in contending that we can endorse Rawls's political philosophy without accepting the machinery of the original position.⁵⁵

How would contractualist liberalism arrive at determinate principles from the abstract commitment to persons as free equals *without* the original position? Rawls himself offers us an approach. While he never abandons the ideal of the original position, his later work demonstrates that he thinks norms of justice can be identified simply by assessing what it is all reasonable persons could agree upon. Let us consider this approach in greater detail, as I will refer to it later. While reasonableness originally appears in the Rawlsian canon as a feature of a decision procedure—modelled by the informational deprivation suffered by parties to the original position⁵⁶—in *Political Liberalism* Rawls casts it as a virtue of citizens, corresponding to the moral power of possessing a “sense of justice.”⁵⁷ On the view of reasonableness as a virtue, it is characterised by two features: firstly, “the willingness to propose fair terms of social cooperation and to abide by them provided others do,”⁵⁸ thereby reflecting “a desire for its own sake a social world in which [citizens], as free and equal, can cooperate with others on terms all can accept;”⁵⁹ and secondly, “the willingness to recognize the burdens of judgment⁶⁰ and to accept their consequences for the use of public reason in directing

⁵⁵ Brian Barry, *Justice as Impartiality* (New York: Oxford University Press, 1995). p. 67ff.

⁵⁶ *A Theory of Justice* does not cast reasonableness in this role, but Rawls clarifies that reasonableness is indeed modelled by the veil of ignorance in his Dewey Lectures. Rawls, “Kantian Constructivism in Moral Theory,” p. 317.

⁵⁷ Rawls, *Political Liberalism*, p. 19.

⁵⁸ This stipulation does not imply that the principles of justice are morally optional; it relates to the significance of *assurance*, which I discuss in Chapters Two and Five.

⁵⁹ *Ibid.*, pp. 54, 50.

⁶⁰ For criticism that denies that the burdens of judgement are necessary for the account, see Leif Wenar, “Political Liberalism: An Internal Critique,” *Ethics* 106, 1 (1995). For criticism of Wenar's proposal, see Larry Krasnoff, “Consensus, Stability, and Normativity in Rawls's Political Liberalism,” *The Journal of Philosophy* 95, 6 (1998): 269-292, p. 279n.

the legitimate exercise of political power⁶¹ in a constitutional regime.”⁶² Reasonable persons are those who effectively constrain their *rational* pursuit of their good in the ways enjoined by reflection on what the idea of reasonableness demands. We can summarise this account of reasonableness by saying that reasonable persons are committed to respecting the moral status of others as free and equal by regulating their conduct by principles that are acceptable to others *as* free and equal, given the burdens of judgement.⁶³

It is widely remarked that, in contractualist theories, there is no *agreement* that takes place. If there is one best argument for what the correct principles are, then surely it should grip the minds of *every* morally motivated participant. Given the stipulation of perfect reasoning on the part of the parties, this objection is sound.⁶⁴ While we *aspire* to achieve real-world agreement on the moral norms we take to be correct—because we aspire to inhabit a world in which everyone acts justly—the question of whether a proposed norm is a morally wise or unwise proposal for regulating our collective affairs does not depend on whether it actually commands assent. Norms of justice are “certain fundamental norms of conduct, possessing an objective validity, [which] can serve as a basis for judging others, whether or not they meet with agreement.”⁶⁵ Nevertheless, it is

⁶¹ Given what I said in the previous section about coercion, we can read “power” in this statement to refer to *normative power*: the right to alter others’ normative situations by creating fresh reasons for action. See Joseph Raz, *Practical Reason and Norms* (Oxford: Oxford University Press, 1999), p. 98. I treat this issue of authority briefly at the chapter’s end.

⁶² Rawls, *Political Liberalism*, p. 54.

⁶³ Brian Barry describes in greater detail a method at arriving at determinate principles without the original position, drawing on T. M. Scanlon’s contractualist theory. I shall have occasion to discuss Barry’s method in Chapter Three. Barry, *Justice as Impartiality*, pp. 67–72.

⁶⁴ Charles Larmore, “The Moral Basis of Political Liberalism,” p. 616. Here Larmore is criticising Habermas’s view, which builds discourse into what it means to arrive at justified conclusions, such that justification cannot proceed apart from discourse. This is significantly different from the Rawlsian/Scanlonian claim that I endorse, and that Larmore himself endorses, contending that discourse is *epistemically beneficial* (indeed, given human limitations, *practically indispensable*) to arriving at reasonable judgements about what we have reason to do.

⁶⁵ *Ibid.*, p. 620.

important to see that while actual agreement is dispensable for justification, the *idea* of a fair agreement is helpful in organising our analyses about what is mutually justifiable, helping us to imagine the conditions in which we could establish that a principle was justifiable to all. Moreover, it is worth adding that insofar as we know that we will always make mistakes in our reasoning, the task of seeking to understand what lies behind *actual, real-world* disagreements about justice can help us understand where we or others might have gone wrong in our reasoning—an insight I shall develop to a significantly greater extent in Chapter Three.⁶⁶

§3. Reasonable Disagreement the Role of Democracy

No matter whether we employ the device of the original position to generate specific principles, or if we simply ask what principles could command agreement among idealised versions of ourselves when reasoning at our best, reasonable citizens are bound to experience *reasonable disagreement about justice*: disagreements about what is demanded by their shared ideal of respecting one another as free and equal. Citizens disagree about whether the principle of free speech protects hate speech; they disagree about whether religious freedom demands that minorities burdened by generally applicable laws be granted exemptions; they disagree about distributive justice and the relation between personal responsibility and work; they disagree about the proper rates of economic

⁶⁶ This description is not to suggest that the reasoning is wholly *deductive*. While Rawls writes that “[w]e should strive for a kind of moral geometry with all the rigor which this name connotes,” he concedes that his own work tends to “fall far short of this, since it is highly intuitive throughout.” *A Theory of Justice*, p. 105. Rawls uses the method of *reflective equilibrium* to move back and forth between abstract principles and considered judgements in particular cases. Rawls’s aspiration is that we *construct* the deductive theory of reasoning through reflective equilibrium; “we may think of the interpretation of the original position that I shall present as the result of such a hypothetical course of reflection.” *A Theory of Justice*, p. 18. What that means, however, is that if the contractualist procedure were to yield the conclusion that slavery is morally permissible, we would not be committed to the implausible conclusion that we should revise our judgement about slavery’s moral permissibility; we should go back and rethink the procedure.

growth and the demands of intergenerational justice; they disagree about bioethics and the proper role of government in scientific research; and they disagree about the character and purposes of their foreign policy. The claim that these disagreements are *reasonable* is the claim that good arguments can be adduced, on the basis of the underlying political liberal ideal, to support various positions on many sides of these debates. Citizens who “share a common human reason” and hold “similar powers of thought and judgment”—who are equally equipped to “draw inferences, weigh evidence, and balance competing considerations”—nevertheless arrive at different positions on these issues.⁶⁷ This is not to say that there is no right answer to the questions about which such citizens disagree. It is simply to say that the questions are extremely difficult to answer, owing to *the burdens of judgement* I introduced earlier. These burdens include the fact of mixed evidence on many issues, the role of intuition in assigning weight to different considerations, the indeterminacy and vagueness of different concepts, and the inevitable role of experience in crafting our commitments.⁶⁸

Nevertheless reasonable disagreements about justice are not like reasonable disagreements about the good life. The fact that we reasonably disagree about the good is part of our normative self-understanding: free equals permit each other, out of respect, to find their own way on some of life’s most complicated questions, and we do not endeavour (in the political realm) to eliminate such disagreements. But reasonable disagreements about justice are crucially disanalogous, and thereby pose a distinctive problem. From the fact that we disagree about tax policy, it cannot follow that we ought not to have a tax policy. Letting our reasonable disagreements about justice serve to

⁶⁷ Rawls, *Political Liberalism*, p. 55.

⁶⁸ *Ibid.*, p. 56.

paralyse our decision-making would involve giving up on the very thing we are disagreeing *about*: the achievement of justice. What is the solution?

Jonathan Quong has observed that reasonable disagreements about justice do not have the same “foundational” character as reasonable disagreements about the good or other sectarian concerns implicated in citizens’ respective comprehensive convictions.⁶⁹ Firstly, reasonable citizens *agree* about a great deal of what justice requires. Certain proposed policies—banning racial minorities from expressing political views, depriving those born poor of opportunities to work for a decent standard of living, permitting workplace discrimination against women—are deemed unjust by *all* reasonable citizens. Secondly, for those proposed policies about which we reasonably disagree—such as tax policy—we agree that *some* reasonable solution to the question is necessary.⁷⁰ As Quong writes:

Suppose, for example, that justice requires the state to enforce one regime regarding property rights from the set (P1, P2, P3, P4), but citizens are evenly divided regarding which regime from the set is in fact required by justice. Further suppose that a world where the state fails to enact any of the property rights regimes (i.e. it does not enact any of P1-4) will be far less just than a world where *any one* of P1-4 are enforced.⁷¹

Quong’s contention is that reasonable citizens prefer *some* reasonable solution from the set to be selected as the official property scheme, even if their preferred solution is not selected.⁷² Indeed, they see themselves as *morally required by justice* to settle for some

⁶⁹ Quong, *Liberalism Without Perfection*, p. 193.

⁷⁰ This relates to what Jeremy Waldron has called “the circumstances of politics.” I shall have occasion to take up this issue in greater detail in Chapter Four. Waldron, *Law and Disagreement* (New York: Oxford University Press, 1999), p. 102.

⁷¹ Quong, *Liberalism Without Perfection*, p. 133.

⁷² Rawls presents three criteria for what qualifies a proposed conception of justice as a *reasonable* conception within this family: that it provides basic liberties, that it establishes equality of opportunity, and that it guarantees the satisfaction of material basic needs. This, of course, concerns reasonable disagreement among conceptions. When it comes to particular interpretations of even a single conception, there is bound to be intense and reasonable disagreement, especially once contentious empirical and social-scientific information serves as important premises of arguments for particular public policy conclusions.

reasonable solution in the face of disagreement; otherwise, *ex hypothesi*, their society would become “far less just,” a result that reasonable citizens regard as morally intolerable.

But a problem then arises: how should citizens choose among reasonable proposals? It will not do for one citizen to start commanding the others around, insisting that *her* preferred conception of justice be implemented. Because we *all* possess the first moral power—a capacity for justice—we must arrive at our solution in a manner that respects all involved: a manner that will deliver provisional conclusions to our disagreements that stand, authoritatively, in the name of all. *Democracy*, of course, is the familiar answer. But why? ⁷³ Eric Beerbohm has recently advanced a powerful interpretation of democracy according to which it secures its value precisely by placing us in a relationship of mutual accountability over one another:

My suggestion is that we embed more directly the contractualist insight that standing in a relation of mutual accountability has immediate and irreducible value. On this approach, democracy is fundamentally a way of relating to others as simultaneously coauthors and cosubjects.⁷⁴

My interpretation of this view’s distinctive appeal is this: our status in a *democracy* as co-authors and co-subjects reflects, *mutatis mutandis*, our *moral* status as both responsible for treating one another in certain ways and also entitled to be treated in certain ways. If justice is about conducting our civic relationships in accordance with certain standards that honour the status of all as free and equal, democracy is what empowers us actually to *undertake* those relationships. We decide decisions democratically—listening to each other and subsequently taking a vote that affirms the status of each citizen as an equal participant whose voice matters—not because democracy happens to secure optimal

Rawls, “The Idea of Public Reason Revisited,” in *Political Liberalism*, p. 450.

⁷³ There are many views on why democracy secures its value, and I will have occasion to examine some of them in Chapters Three and Four.

⁷⁴ Beerbohm, *In Our Name*, p. 37.

outcomes for citizens' well-being, but because it enables us to exercise our first moral power.

The value of this approach to democracy, Beerbohm rightly notes, is that it defies the familiar dichotomy between intrinsic defences of democracy, according to which democracy secures its value because of some property (e.g., fairness⁷⁵) that its procedure enjoys, and instrumental defences, according to which democracy is simply some kind of “technology”⁷⁶ to be evaluated at arm’s length for its prospective benefits to the human beings it serves.⁷⁷ But neither strikes us as an accurate understanding of why we value democracy. Take the idea that democracy has intrinsic value. This suggests that when atrocities are perpetrated democratically, our reaction ought to be: “That was awful—but at least there was *some* value to it, given it was democratic.” Beerbohm notes that we are likely to experience the opposite response; the *democratic* perpetration of an injustice makes it *worse*, reflecting not the whims of one authoritarian madman but the coordinated effort of numerous people to wrong others.⁷⁸ But neither does the instrumental defence of democracy fit our phenomenology. Our commitment to democracy, at its core, is not provisional—awaiting empirical investigations into alternate forms of institutional design.⁷⁹ This is not simply because the undergraduate fantasy of benevolent dictatorship is implausible. Rather, we believe in democracy because it matters to us that *we* treat

⁷⁵ For an argument as to why procedural fairness is always a *pro tanto* good, see John Broome, “Fairness,” *Proceedings of the Aristotelian Society* 91, 1 (1991): 87-101.

⁷⁶ Beerbohm, *In Our Name*, p. 37.

⁷⁷ Richard Arneson, “Democracy is Not Intrinsically Just,” in *Justice and Democracy: Essays for Brian Barry* edited by Keith Dowding, Robert E. Goodin and Carol Pateman, Cambridge: Cambridge University Press, 2004.

⁷⁸ Beerbohm, *In Our Name*, p. 36.

⁷⁹ This is not to say that there is not room for considerable variation in how democracy is structured and what the right division of labour ought to be. I shall assume for the purposes of this dissertation that democracy involves majority rule, voting for officials on the basis, *inter alia*, of their policy commitments, as well as various substantive referenda.

others as free and equal; the core task of specifying the fundamental terms by which we are to live and maintaining them is not to be outsourced.

The fact that a decision's democratic pedigree increases our moral horror in instances of injustice shows that the mere status of a government as democratic is woefully insufficient for it to have *legitimate authority*, which for my purposes here I shall define simply as *the moral power to determine which reasonable courses of action citizens ought to pursue*.⁸⁰ Rather, on the view I shall be assuming, a democracy possesses legitimate authority only if it has a strong tendency to enact *reasonable* legislation, since only then can it possibly live up to the definition I have just offered of what a legitimate authority *is*.⁸¹

Note the compatibility of this account with Rawls's principle of legitimacy:

[O]ur exercise of political power is fully proper only when it is exercised in accordance with a constitution, the essentials of which all citizens as free and equal may reasonably be expected to endorse in the light of principles and ideals acceptable to their common human reason.⁸²

On my interpretation, "power" in that phrase need not be interpreted to refer to coercive power, but simply the moral power to create reasons for action; when the state declares one reasonable policy over another as the provisional position of the democratic community on that subject matter, citizens have strong moral reasons to act accordingly that they previously lacked. Moreover, if Beerbohm is right about democracy's value, then free and equal citizens would *only* design a constitution that was democratic in character. Thus the two conditions I shall stipulate of what makes a law *legitimate*—that it

⁸⁰ For the most influential contemporary work on authority, see Joseph Raz, *The Morality of Freedom*, pp. 38-109. I do not propose to treat the theory of authority here, as it does not arise in the chapters to come and would take me too far afield. I rely on Rawls's position that we have a natural duty of justice to support and comply with just institutions. Rawls, *A Theory of Justice*, p. 99. See also Jeremy Waldron, "Special Ties and Natural Duties," *Philosophy & Public Affairs* 22, 1 (1993): 3-30.

⁸¹ This is compatible with the Razian formulation offered in Quong, *Liberalism Without Perfection*, p. 130.

⁸² Rawls, *Political Liberalism*, p. 137.

be *reasonable*, and that it be *democratically enacted*—both flow from Rawls’s own principle of legitimacy.⁸³

I have nearly completed my presentation of the various elements of political liberalism to orient the chapters to come. But there is one more question that I must address before proceeding. That is the question of what constitutes citizens’ common ground. In *A Theory of Justice*, Rawls argues that citizens of his ideal society would share the same conception of justice—his own *justice as fairness*. Over time, however, Rawls sees that the fact of reasonable disagreement about justice makes it likely that citizens would not affirm the same conception of justice, but rather “a class of liberal conceptions.”⁸⁴ While he never wholly updates his understanding of a well-ordered society—a society in which citizens’ shared convictions about justice are successfully manifested in their relations—Jonathan Quong has offered the plausible argument that the common ground of citizens is best conceived as the normative materials that all reasonable conceptions hold in common.⁸⁵ The elements discussed in this chapter are best conceived as the content of that common ground. In the chapters to come, I shall refer to this common ground as *the commitment to justice*: the commitment to a social world in which citizens live together as free and equal by regulating their relations by democratically specified, reasonable terms of cooperation.

⁸³ Note that on this view, a law enacted by a legitimate state is not therefore legitimate; it must also be reasonable. As Rawls writes, “[a]t some point, the injustice of the outcomes of a legitimate democratic procedure corrupts its legitimacy...” Rawls, “Reply to Habermas,” in *Political Liberalism*, p. 428. Note also that this rests in conflict with Weithman’s interpretation of Rawlsian legitimacy, according to which we only come to understand what is legitimate or not once we already have a determinate, shared conception of justice on hand. I reject Weithman’s view as incompatible with the fact of reasonable disagreement about justice and the role that the concept of legitimacy properly plays in our practical reasoning. Weithman, *Why Political Liberalism*, pp. 312-21.

⁸⁴ Rawls, *Political Liberalism*, p. 164.

⁸⁵ Quong, *Liberalism Without Perfection*, p. 185.

Conclusion

To have a commitment to justice is to view others in accordance with a particular understanding of human moral personality that regards persons as *free and equal*—as responsible human agents of equal moral worth who ought to be respected in framing, revising, and pursuing their chosen conception of a good life; to identify norms of conduct that are mutually acceptable among persons so understood; to settle reasonable disputes about those norms *democratically*; and to regulate one's conduct accordingly. Our normative ideal is now before us. Now imagine a society that has realised it. Could it last? Would it last? The answers to these questions matter. The next chapter explains why.

CHAPTER TWO

The Problem of Stability

I have identified the foundational elements of a commitment to justice. These foundational elements constitute the moral core of liberal democracy. While a society whose citizens and institutions are united around that public conception may not be perfectly just—we will enduringly disagree about which standards' satisfaction warrants *that* designation—we can regard it as a well-ordered liberal democracy. It is a society in which citizens do not express fundamental regrets about the character of their social relations, and in which they stand together in affirmation of their shared liberal values.

The public conception of a well-ordered liberal democracy that I have offered, I submit, rests in concord with our considered convictions about the standards of treatment that we owe to each other as free and equal citizens. Let us take that proposition for granted. The long journey of this chapter begins with a question: is this proposition sufficient to regard the public conception as justified—as morally binding? Common sense suggests it is. What else *could* be relevant to determining the fundamental standards of our public morality beyond our reflectively endorsed convictions about the content of those very standards? John Rawls has an answer. Before a public conception is justified, Rawls contends, it must be shown to be *stable*: as generative of its own “freely willed” support—support, that is, from citizens' sense of justice—in the societies it regulates, given facts of human nature. If a public conception fails to meet this *stability test*, then it is *defective* as a public conception. It must be rejected.

In this chapter my aim is to defend the contention that a liberal democracy's governing public conception is radically incomplete without an account of its own stability, but to do so in a way that reconceives—and extends—Rawls's particular understanding of stability's significance. My overarching thesis is that while Rawls is right

to think that freely willed stability matters, he is wrong both about the precise way in which it matters, and about what, exactly, it can involve—leading him to misidentify what follows from stability’s significance for the practice of citizenship.

Specifically, I shall argue that Rawls is wrong to assign freely willed stability justificatory significance; it defies what it is to hold a moral commitment to abandon or dilute it upon discovering that fallible human beings are unlikely to live up to its demands. Nevertheless I shall argue that freely willed stability retains profound normative significance for two reasons, both flowing from what a commitment to justice *is*. Firstly, the commitment to justice is not merely a commitment to some distribution of freedom and resources, but to an underlying ideal of respectful civic relationships, themselves constituted by certain attitudes and their appropriately motivated manifestation in reliable patterns of conduct. For social relations to *be* fully just, they must be freely willed. Secondly, those who hold a commitment to justice should be seriously disturbed by the empirical discovery that the successful operation of their sense of justice might face potentially overwhelming resistance from human nature. They should—if they are committed to justice—resolve to *do* something about it. But *because* justice concerns the quality of their relations, their resolution cannot merely be to enact repressive penal sanctions, terrorising people into doing their duty. Rather, they should resolve to *fortify* their sense of justice against countervailing threats, designing and submitting to civic practices that increase the likelihood that they will accurately identify and then freely comply with their fundamental duties of justice. If we worry—as the citizens of the well-ordered society should—that our sense of justice may face overwhelming resistance from elements of our nature unless steps are taken to fortify it, the appropriate course of action is to pursue precisely such fortification.

In Section 1, I review the idea of stability as it appears in Rawls's work. In Section 2, in jointly exegetical and critical spirit, I canvass some possible justifications for regarding stability as justificatorily significant in the way Rawls does, demonstrating why each is defective. In Section 3, I explain the way that concerns of stability nevertheless retain normative significance, arguing that a commitment to justice properly involves not only an aspiration to treat people as free and equal, but also a conscientious mission to increase the likelihood of that aspiration's success by designing and submitting to certain "fortificational measures." In Section 4, I offer greater detail on the content and structure of my proposal, and in Section 5, I take up some concerns about these measures' justification and limits.

A final preliminary word is necessary. Rawls speaks mostly in *A Theory of Justice* and *Political Liberalism* about the stability of a society's commitment to a *particular conception of justice*. As we saw in Chapter One, the proper focus of consensus is considerably thinner, and more fundamental, in light of reasonable disagreement about justice. Given that Rawls himself recognises this point, it is plausible to interpret Rawls as suggesting that the stability test applies to *whatever* the focus of citizens' consensus properly is. But because I engage here with so much of Rawls's text, and he continually refers to the stability of *a conception of justice*, I will not continually adjust or correct for this in my prose or quotations. I will sometimes use the suitably ambiguous phrase "public conception." Note that for readers who *do* think that a determinate conception of justice is likely to command the reasonable assent of all in a well-ordered society, as Rawls initially hopes, my argument here still bites; it applies to *whatever* the central moral ideal that organises our political life might be.

§1. Rawls's Stability Argument

1.1. Stability from *A Theory of Justice* to *Political Liberalism*

According to Rawls, there are two stages to justifying a public conception: firstly, the identification of a public conception of justice, and secondly, the test of its stability. Stage 1 familiarly proceeds by imagining parties to the original position assessing which principles increase their bundle of primary goods, though we need not imagine Stage 1 using this heuristic device. Rather, we can simply imagine it as the stage in which we ask: *which public conception best expresses our commitment to the status of others as free equals, in light of our considered convictions about what respect for that status demands?* Only once we have answered this question, in whatever form asked, do we proceed to the matter of stability. While most philosophical analysis of Rawls has focused on the first question, this chapter attends to the second.

For Rawls stability is “an attribute of moral principles.”¹ Specifically:

An important feature of a conception of justice is that it should generate its own support. Its principles should be such that when they are embodied in the basic structure of society men tend to acquire the corresponding sense of justice and develop a desire to act in accordance with its principles. In this case a conception of justice is stable.²

Rawls clarifies just *how* important this feature of stability is, in both early and late work:

However attractive a conception of justice might be on other grounds, it is *seriously defective* if the principles of moral psychology are such that it fails to engender in human beings the *requisite* desire to act upon it.³

[S]tability is imposed as a *condition* on a reasonable conception of political justice: such a conception must be able to generate its own *sufficiently strong* supporting sense of justice.⁴

¹ George Klosko, “Rawls’s Argument from Political Stability,” *Columbia Law Review* 94, 6 (1994): 1882-1897, p. 1883.

² Rawls, *A Theory of Justice*, p. 119.

³ *Ibid.*, p. 398. Emphasis added.

⁴ Rawls, *Justice as Fairness*, edited by Erin Kelly (Cambridge: Harvard University Press, 2001), p. 89. Emphasis added.

There are numerous ambiguities in these formulations which I shall discuss in due course. For now, I want to press on what it means for a public conception to *generate* its own support. Rawls believes that there are two elements to this: firstly, whether citizens would *acquire* a sense of justice; and secondly, whether citizens would assign that sense of justice *regulative primacy* over their practical reasoning. Call these the issues of *acquisition* and *maintenance*.

By saying that a public conception must generate its own support, what does Rawls mean? We can find a clue in a variation Rawls offers on the above definition, where he says that a conception of justice is stable if “inevitable deviations from justice are effectively corrected or held within tolerable bounds *by forces within the system*.”⁵ Now the “inevitable deviations” to which he refers here are *not* deviations from justice due to malicious or ignorant non-compliance. He does not regard *those* as inevitable. Rather, he is explicitly referring to deviations from justice generated by social change—as, for example, when spontaneous changes in the economy mean that the current tax system no longer happens to comport with the relevant principle of distributive justice and needs to be updated.⁶ For our purposes, what matters is the phrase “forces within the system.” Now we might think this refers to any forces *within* the just society itself; on this view, a society’s justice is stable if it does not need foreign actors to come help re-adjust it after deviations from justice. But this is too capacious an interpretation. Rawls has something more specific in mind. In a paper published shortly before *A Theory of Justice*, Rawls writes: “a conception of justice is stable *if the institutions which satisfy it* tend to generate their own support...”⁷ Paul Weithman interprets this passage to mean that the

⁵ Rawls, *A Theory of Justice*, p. 401. Emphasis added.

⁶ *Ibid.*

⁷ Rawls, “Distributive Justice: Some Addenda,” in *Collected Papers*, p. 171. Emphasis added. Paul Weithman flags this passage in *Why Political Liberalism*, p. 45.

stability-providing institutions are simply “the basic institutions that would be established to implement the principles of distributive justice.”⁸ This is important, as it reflects an element of Rawls’s understanding of stability that will be central to my analysis. Stability is traceable to *justice-delivering institutions alone*. As Weithman writes:

Some parts of some societies’ governing apparatus—such as repressive penal institutions—might be established precisely because the institutions which satisfy the conception of distributive justice in question are *incapable* of generating their own support.⁹

But such *external practices*—designed precisely to secure compliance—have no place in Rawls’s view: justice-delivering institutions must themselves be adequate to cultivate citizens’ *acquisition* and *maintenance* of their sense of justice. If “repressive penal institutions” are needed to secure compliance with a conception, given human nature, it is not stable on Rawls’s understanding. The stability of a public conception is thus a matter of citizens’ freely willed compliance with it, *traceable to the fact that they are moved by their sense of justice to regard it as reasonable or correct and thus worthy of affirmation*. Only then is there *genuine* stability: stability “for the right reasons.”¹⁰

So let us now turn directly to the issues of *acquisition* and *maintenance*. Note that each actually embodies two distinct concerns. The concern with acquisition is that of whether citizens will acquire a disposition to act justly *and* with whether members of a well-ordered liberal democracy will come to affirm *liberal convictions* to which their sense of justice will be indexed. The concern with maintenance is a concern with whether members of the well-ordered liberal democracy will *maintain* their sense of justice—continually granting it *regulative primacy* over their practical reasoning—*and* whether they sustain the liberal convictions to which that sense of justice is indexed, continuing to

⁸ *Ibid.*, p. 46.

⁹ *Ibid.*, p. 45.

¹⁰ Rawls, *Political Liberalism*, p. xxxix.

apply them intelligently to fresh issues and concerns. Rawls's discussion conceals the epistemic dimensions of acquisition and maintenance, appearing to assume that if *any* sense of justice is acquired and maintained in a liberal society, it is bound to be a liberal one. Because liberal values are ripe for misinterpretation, especially in forever changing times, I think this assumption is unwarranted. Accordingly, my discussion later on will pay attention to these epistemic dimensions of acquisition and maintenance.¹¹

Rawls's argument in *A Theory of Justice* for acquisition—that people raised in a well-ordered society will acquire a sense of justice—centres on a theory of moral development. According to Rawls, human beings have a naturally selected tendency toward reciprocity that socialisation under well-ordered liberal institutions calls forth. Through experiencing the love of parents, children learn to reciprocate love, and the tendency for reciprocity is then refined and cultivated as children participate in different settings with different roles and attendant expectations they must fulfil. As children come to learn what their society's public conception is—and its status as *public* ensures that they will—they will come to see the expectations that apply to them as a citizen of the well-ordered society, and will *want* to act on these expectations when each “recognizes

¹¹ Rawls himself understands the sense of justice as involving an epistemic component. He defines the sense of justice as “the capacity to understand, to apply, and to act from the public conception of justice which characterizes the fair terms of social cooperation.” He adds that if people are socialised into a liberal conception, “a sense of justice also expresses a willingness, if not the desire, to act in relation to others on terms that they also can publicly endorse.” I propose that we modify our understanding of the sense of justice to include the capacity to *reason* about the relative merits of different conceptions of justice, in addition to the other elements Rawls offers; this accommodates the fact of reasonable disagreement about justice. Further detail on what the sense of justice is takes us into controversial debates in moral psychology; I will have something to say about those debates later in this chapter, as well as in Chapter Five. See Rawls, *Political Liberalism*, p. 19. It is important to distinguish the sense of justice as a moral power from the sense of justice as colloquially used to refer to the emotional gratification one incurs when justice transpires, or the outrage one experiences when injustice transpires. For reflection on the latter with some reflections on its connection to the former, see Markus Dirk Dubber, *The Sense of Justice: Empathy in Law and Punishment* (New York: New York University Press, 2006), as well as the classic treatment of the topic in Judith Shklar, *The Faces of Injustice* (New Haven: Yale University Press, 1990).

that he and those for whom he cares are the beneficiaries of these arrangements.”¹² Moreover, the educative effects of living under just institutions—in which “[t]hey are presented with a way of regarding themselves that otherwise they would most likely never have been able to entertain”¹³—will ensure that children grow up to think of themselves and others as free as equal. Rawls never abandoned this argument for the acquisition of the sense of justice.¹⁴ More controversial, and subject to extensive revision, was the argument for *maintenance* that Rawls offered in *A Theory of Justice*.

The argument for maintenance rests on a proposition that *even if* citizens have acquired a sense of justice, they may come to *resent* it if they experience it as continually contrary to their interests. If justice and goodness are not “congruent,” citizens may opt to become the sort of people who decide whether to honour their sense of justice case-by-case—which, of course, is not to honour it at all, given that the sense of justice is, by its nature, “executive and regulative,” and thus not “on the same footing” with other “natural inclinations.”¹⁵ In order to be sure that his conception of justice does not lead people to become like that, Rawls tries to show that citizens *would* find justice congruent with their own good. And he does so precisely by sketching a partial comprehensive doctrine that *A Theory of Justice* avows would be shared by most members of the well-ordered society.

I cannot do justice to Rawls’s elaborate argument on this topic here; it occupies two of the most complex (and understudied) chapters of *A Theory of Justice*. I will simply note a few features of the conception of good Rawls imagines that members of the well-

¹² Rawls, *A Theory of Justice*, pp. 429-431.

¹³ Rawls, “Kantian Constructivism in Moral Theory,” p. 340.

¹⁴ He downplays its empirical elements in *Political Liberalism*, emphasising its role as part of a normative conception of the person; but he explicitly says he continues to accept the arguments on acquisition from *Theory* when he addresses the issue in *Justice as Fairness*, p. 184.

¹⁵ Rawls, “Kantian Constructivism in Moral Theory,” p. 320.

ordered society would come to hold. Firstly, the partial conception includes such commitments as citizens' "desire to express their nature as free and equal moral persons."¹⁶ And since "[t]he desire to act justly and the desire to express our nature as free moral persons turn out to specify what is practically speaking the same desire," one's self-interested desire to express her nature will embolden her commitment to justice.¹⁷ (Because of its emphasis on autonomy, this first prong is often termed "the Kantian congruence argument."¹⁸) Secondly, under the heading of the so-called Aristotelian Principle, Rawls argues that citizens will "enjoy the exercise of their realized capacities (their innate or trained abilities), and this enjoyment increases the more the capacity is realized, or the greater its complexity."¹⁹ He continues: "As we witness the exercise of well-trained abilities by others, these displays are enjoyed by us and arouse a desire that we should be able to do the same things ourselves. We want to be like those persons who can exercise the abilities that we find latent in our nature."²⁰ Citizens will not only enjoy exercising their sense of justice; they will revel in their enjoyment of a public culture in which the autonomously developed excellences and talents of all are freely supported and celebrated. Thirdly, citizens will have desires for love and friendship, and their ties with friends will naturally extend to a widespread civic friendship with one's fellow citizens.²¹ Fourthly, citizens will want to pursue honest relations with others, seeking to avoid the costs generated by "a systematic course of deception and

¹⁶ Rawls, *A Theory of Justice*, p. 462.

¹⁷ "[B]y acting from these principles [of justice as fairness] persons are acting autonomously: they are acting from principles that they would acknowledge under conditions that best express their nature as free and equal rational beings" and that "acting from this precedence expresses our freedom from contingency and happenstance." Rawls, *A Theory of Justice*, pp. 500, 452, 503.

¹⁸ Weithman, *Why Political Liberalism*, p. 184ff.

¹⁹ Rawls, *A Theory of Justice*, p. 374.

²⁰ *Ibid.*, pp. 375–76.

²¹ *Ibid.*, p. 500.

hypocrisy”.²² There are more elements to Rawls’s account, but we need not bother with them here.²³

Famously, Rawls became intensely dissatisfied with the stability argument in *Theory*, recasting it extensively in *Political Liberalism*. Now, as Paul Weithman has rightly pointed out, many scholars interpret Rawls’s “political turn”—implicitly, anyway—as principally motivated *not* by an anxiety about stability, but rather by an anxiety about the theory’s fundamental justification—its justificatory starting point. On this interpretation of Rawls’s political turn, termed the “Public Basis View” by Weithman, the reason Rawls recast his argument is because he came to realise that he had predicated the original position in *A Theory of Justice* on a sectarian account of the person derived from Kant’s metaphysics.²⁴ He realised that such a theory was therefore *illiberal*, since free and equal citizens reasonably disagree about precisely such sectarian questions; accordingly, on this view, Rawls quietly dropped his theory’s Kantian foundations, insisting instead that his theory would be *silent* on its deep justification and that citizens could come to endorse liberal principles through their own paths, coming to form an “overlapping consensus” on liberal values.²⁵ Rather than upholding justice because of its role in enabling a Kantian escape from the heteronomous contingencies of the phenomenal realm, or a Millian celebration of diverse excellence, citizens would support the ideal of seeing others as free equals for distinct reasons internal to the comprehensive doctrine of each. So described,

²² *Ibid.*, p. 499.

²³ I am indebted to the instructive discussions in Weithman, *Why Political Liberalism*, pp. 97-121, and also Gerald Gaus, “The Turn to a Political Liberalism,” in *A Companion to Rawls*, edited by David Reidy and Jon Mandle (New York: Wiley-Blackwell, forthcoming), and Samuel Freeman, “Congruence and the Good of Justice,” in *Justice and the Social Contract: Essays on Rawlsian Political Philosophy* (New York: Oxford University Press, 2007).

²⁴ Weithman, *Why Political Liberalism*, p. 17.

²⁵ Paul Weithman accuses both Charles Larmore and Bruce Ackerman of adhering to the Public Basis View in *Why Political Liberalism*, p. 31n. See Larmore, “Political Liberalism,” pp. 345-46, and Ackerman, “Political Liberalisms,” *The Journal of Philosophy* 91 (1994): 364-86, p. 365.

I believe that the Public Basis View accurately captures one indispensably important element of the transition from *A Theory of Justice* to Rawls's *Political Liberalism*. But Weithman is right that it cannot be the whole story.²⁶ For, Rawls explicitly says that the change in *Political Liberalism* results from changes to the *stability* argument of *A Theory of Justice*—Part III—which does not concern the issue of the theory's justificatory starting point; it concerns the issue of the theory's tendency to endure, given facts of human nature.²⁷ The Public Basis View, however, suggests the changes concern Parts I and II.

So consider the second mainstream interpretation of Rawls's political turn—what we can call the Empirical Prediction View. On this view, Rawls's political turn was motivated by an *empirical* realisation that the conception of the good he imputes to all in *A Theory of Justice* as part of his account of stability—with all the desires that I sketched a moment ago—is *unlikely* to be freely realised. Because of the troublesome fact of philosophical and religious diversity, Rawls came to see that he had to *dilute* his view in order for it to gain widespread assent. This account has its many adherents, though as will become clear throughout discussion in this chapter, I believe it is implausible, recasting the project of public justification as an empirical search for overlaps in people's existing belief sets.

The best exegetical interpretation of Rawls's political turn is what I will call the Reasonable Pluralism View. On this view, Rawls's political turn was indeed motivated by stability, but not (*pace* the Empirical Prediction View) as a matter of contingent sociological prediction. Rawls's discovery between *A Theory of Justice* and *Political Liberalism* was not the empirical discovery that members of the well-ordered society

²⁶ It is not clear that Weithman even thinks it is part of the story *at all*, given the fact that Rawls already builds justice as fairness in *A Theory of Justice* out of convictions that he says “we do in fact accept,” through the method reflective equilibrium, and talks of Kant rarely in the early sections. Rawls, *A Theory of Justice*, pp. 19, 221ff.

²⁷ Rawls, *Political Liberalism*, p. xvi.

would not *happen* to affirm the same (partially²⁸) Kantian comprehensive doctrine. Rather, he saw that in a society that affirms a view of citizens as free and equal, and encourages their exercise of freedom of expression and association, the doctrine of the good sketched in *Theory* would almost certainly be rejected by many conscientiously reasoning citizens. “[I]he fact of reasonable pluralism [is] itself the outcome of the free exercise of free human reason under conditions of liberty.”²⁹ It is untenable to predicate one’s account of stability on an assumption—ethical homogeneity—that defies one of the defining features of the social world one is endeavouring to stabilise.³⁰

For that reason, Rawls—while upholding the argument for acquisition sketched in *Theory*—abandoned the argument for maintenance. Believing that the socialising pressures of life in liberal societies are sufficiently strong that mainstream comprehensive doctrines would endorse liberal ideals, Rawls altered his argument for why his ideal of justice would generate its own support: because members of a liberal society would each belong to a comprehensive doctrine that affirmed their sense of justice as good for them.

1.2. Ambiguities in Rawls’s stability argument

Notice four ambiguities in the Rawlsian account. The first concerns whether justice is a *relative* or *absolute* measure. Is stability a tiebreaker between two equally promising conceptions of justice—as suggested when Rawls says “other things equal, the preferred

²⁸ It is sometimes suggested—even by Rawls—that *A Theory of Justice* contains a fully comprehensive doctrine. But it is nothing of the sort. There is nothing, after all, in *Theory* that tells people which religion they should use their religious freedom to endorse. Most of the substantive questions about the content of a good life remain utterly unaddressed. One leaves *A Theory of Justice* with no guidance *whatsoever* on whether to become a Christian clarinettist, a Muslim magician, a Buddhist banker, or a Zoroastrian zoologist.

²⁹ *Ibid.*, 144.

³⁰ What I say here mostly comports with Weithman, *Why Political Liberalism*, pp. 234–69.

conception of justice is the most stable one”³¹—or is it a threshold, a binary test that a conception either passes or flunks—as suggested when Rawls says that “stability is imposed as a *condition* on a reasonable conception of political justice: such a conception must be able to generate its own sufficiently strong supporting sense of justice.”³² Different interpretations of stability’s justificatory significance imply different versions of the stability test, so both will be subject to discussion in what follows. But note that *if* we were to vindicate more than one conception of justice as equally morally promising in Stage 1, and all met the some appropriately specified *absolute* stability threshold, they could then be ordinally ranked as more or less stable.

The second ambiguity concerns whether stability is an individual- or population-level judgement.³³ The answer is that it is both. While Rawls is fundamentally concerned about the enduring maintenance of just institutional structures, which lone sociopaths cannot imperil, the achievement of those enduringly just structures is made possible only by the existence of a mass of citizens *whose sense of justice is stabilised against countervailing forces*. The argument for right-good congruence is an argument applying to the stabilisation of each citizen’s internal perceived balance of desires, and to the consequential realisation of a society in which citizens give justice to one another. No individual persons freely willing justice, no freely willed just society.

The third ambiguity is more complicated. What is the philosopher meant to do *after* a conception she establishes as morally promising in Stage 1 is then disqualified in Stage 2 due to stability concerns? Rawls never faces this quandary since he thinks a liberal conception of justice would (or could—depending on the first ambiguity) be stable. There are two possibilities suggested by Rawls. According to the first strategy—the

³¹ Rawls, *A Theory of Justice*, p. 436.

³² Rawls, *Justice as Fairness*, p. 89.

³³ Cf. Gaus, “The Turn to a Political Liberalism,” p. 5.

strategy of concession—the appropriate response is to *dilute* the demands of the initially chosen conception of justice so as to fit the constraints set by human psychological limitations.³⁴ But according to a second strategy—the *strategy of abandonment*—the right response is to *abandon* the task of justifying a conception of justice.³⁵ Rawls’s ambiguity on this matter means that the correct interpretation depends on what the underlying argument for the stability requirement *is*.

The fourth ambiguity’s resolution also depends on specifying the point of the stability requirement. Is stability a matter of *likelihood*, such that a conception of justice is unstable if it is unlikely to be realised or sustained, given facts of human psychology? Rawls sometimes suggests it is,³⁶ making the stability test extraordinarily demanding on the one hand, but also—if the *strategy of concession* turns out to be plausible—insanely concessive, on the other. Perhaps, then, that stability concerns *possibility*: a conception of justice is stable if it *could* generate the conditions of its own support.

But enough preparation: it is now time to find out why any of this should matter for us to be justified in certifying our moral convictions as worthy of our aspiration. All of the possible interpretations I shall consider share similar elements, and they all have some exegetical basis in Rawls, making the value of a comparative exegetical evaluation highly limited and thus only of secondary interest. My main aim is to assess the different arguments on the basis of their plausibility *as arguments* within the general framework of contractualist liberalism.

³⁴ Suggested in *A Theory of Justice*, p. 441.

³⁵ Suggested in *The Law of Peoples*, p. 128.

³⁶ Rawls, *A Theory of Justice*, p. 398.

§2. The Point of Stability

2.1. *The confirmation argument*

According to Thomas E. Hill, Jr., the point of Stage 2 of Rawlsian justification is to *catch mistakes* made in Stage 1 of moral reasoning.³⁷ Hill argues that the point of the stability test “is to show that advocacy and use of justice as fairness...is reasonable and respectful of those whose use of reason leads them to disagree with us deeply.” The stability argument thus “allay[s] doubts that the principles are worthy of actual consent by showing that they could, hypothetically, win (almost) universal agreement if everyone would consider them reasonably.”³⁸ In other words, while we may *think*, in Stage 1, that we have certified a conception of justice as mutually acceptable among free and equal citizens, the fact that the conception of justice would face destabilising resistances from human nature in the societies it regulates is evidence that we have not, in fact, succeeded.³⁹

While there is some highly compressed support in Rawls for this interpretation, this interpretation is exegetically⁴⁰ and philosophically implausible. If the test of stability

³⁷ Thomas E. Hill, Jr., *Pluralism, Respect, and Justice* (Oxford: Oxford University Press, 2000), pp. 237-259.

³⁸ *Ibid.*, p. 252.

³⁹ A similar interpretation is offered in Larry Krasnoff, “Consensus, Stability, and Normativity,” p. 284.

⁴⁰ This footnote addresses the exegetical proposition that Rawls did not merely alter the *content* of the stability test between *A Theory of Justice* and *Political Liberalism*; he also altered the *role* of the stability test. According to this proposition, Rawls *ceased* to be concerned in *Political Liberalism* with the sociological question of the acquisition and maintenance of an effective sense of justice—the chief concern in Part III of *A Theory of Justice*—and instead reconceived the stability test as simply a way of confirming the moral justifiability of his theory. If this were true, it would invalidate my argument in this section (in addition to rendering the overlapping consensus superfluous within Rawls’s view, a fact that impugns this exegetical interpretation). The plausibility of this exegetical proposition appears to be suggested by the following statement:

“[S]eeing whether an overlapping consensus on the political conception is possible *is a way of checking* whether there are sufficient reasons for proposing justice as fairness (or some other reasonable doctrine) which can be sincerely defended before others without criticizing or

were *itself* the test of acceptability to others—the test of whether the conception properly respects the status of all as free and equal—then any distinction between the first and

rejecting their deepest religious and philosophical commitments.” Rawls, “Reply to Habermas,” in *Political Liberalism*, p. 390.

Likewise, Rawls elsewhere comments that “justice as fairness [is] expressly designed to gain the reasoned support of citizens who affirm reasonable although conflicting comprehensive doctrines.” *Ibid.*, p.143. These passages, however, squarely contradict other statements not simply in *A Theory of Justice*, but in *Political Liberalism*, where Rawls writes that the question of stability is one of determining whether “those who grow up under just basic institutions acquire a sense of justice and a reasoned allegiance to those institutions *sufficient to render them stable*,” thereby ensuring that citizens’ sense of justice will be “*strong enough to resist the normal tendencies to injustice*.” *Ibid.*, p. 142. Emphasis added.

Jonathan Quong argues that the overlapping consensus’s distinctive feature is that it concerns the *beliefs* of citizens; that is why we must understand one of its purposes (or, on Quong’s view, its *only* purpose) to be that of serving as the “initial common ground” of citizens’ shared political reasoning (*Liberalism Without Perfection*, pp. 162, 180ff). Yet the whole *point* of asking about whether a conception of justice will be *maintained*—or whether a well-ordered society will, given facts of human nature, degenerate into injustice—is *precisely* because people might choose to *ignore* their beliefs about justice, acquiescing to their darker temptations. This is what *A Theory of Justice* is concerned about, and if we are to believe Rawls (and Weithman’s interpretation of Rawls) that the change to *Political Liberalism* really was motivated by the concern of *stability* as conceptualised in *A Theory of Justice*, we can only conclude that the passages cited at the beginning of this footnote reflect confusion on Rawls’s part: his *real* interest in the overlapping consensus is its *stabilising role in making citizens believe that justice is part of their good*, which will make them less likely to ignore their liberal convictions about justice for selfish reasons.

Now, it is a grand claim to say that a major thinker was confused to include certain passages in his account. But to say otherwise involves imputing to Rawls a seriously implausible position: namely, the position that citizens’ mere *possession* of liberal convictions is “sufficient...to resist the normal tendencies to injustice”—a contention that Rawls *clearly* denies in *A Theory of Justice*, and, I believe, continues to deny in *Political Liberalism* insofar as he remains interested in the question of maintenance (as I am arguing he is). This suggests that Quong’s suggested role for the overlapping consensus—while I think Rawls would have absolutely agreed to be a crucial *element* of the consensus’s role within the theory—nevertheless ignores a vital *additional* purpose that Rawls has for the overlapping consensus: a stabilising role in securing people with an understanding that their sense of justice is *good* for them. (Perhaps Quong ignores this further role because he assumes the sociological reading of the overlapping consensus to be *obviously* justificatorily immaterial; in which case, this chapter aims to vindicate that very assumption.)

Thus I conclude that the best interpretation of Rawls, given his whole project, is that the overlapping consensus really is meant to solve the problem of stability: whether people will be able to acquire and maintain a regulative sense of justice, given facts of human nature and psychology. If we interpret it any other way—for example, in the way suggested by the misleading first passages I quote above—it becomes far too easy to condemn as superfluous, since the test of reasonable acceptance has *already* been carried out in Stage 1. For arguments that it is superfluous, see Brian Barry, as well as Stephen Mulhall and Adam Swift. Barry, “John Rawls and the Search for Stability,” pp. 901ff; Mulhall and Swift, *Liberals and Communitarians*, second edition (Oxford: Blackwell Press, 1996), pp. 240ff. This is not to deny that Rawls is communicating something important when he says: “The kind of stability required of justice as fairness is based, then, on its being a liberal political view...” But what he is communicating here is best understood as the claim that the sort of stability his theory seeks is *freely willed* stability, not stability imposed through repressive schemes of punishment. *Political Liberalism*, p. 143.

second stages of justification (in either *Theory* or *PL*) would disappear. For it to be non-redundant, Stage 2 must ask a different question than Stage 1. And it does: it asks whether the conception “fails to engender in human beings the requisite desire to act upon it,” given “the principles of moral psychology.”⁴¹ This is plainly a different question to whether the principles of justice are mutually acceptable in light of citizens’ status as free and equal. This is *especially* clear when we think of it in reverse: Stage 1’s activities cannot plausibly be re-described as that of determining whether citizens will acquire and assign regulative supremacy to their sense of justice; it is plainly not concerned with such matters of moral psychology.⁴² It is about what principles of justice rest in concord with our convictions about what we owe each other. It is precisely *because* of this stark separation, Rawls insists, that the theory is not plausibly regarded as “political in the wrong way.”⁴³

To be sure, the fact that a conception is not inherently stable *could* be evidence of its potential unjustifiability. If citizens were consistently to rebel against a conception of justice on the grounds that it continually imperils their fundamental interests, such inherent instability could be construed of a sign that the conception fails to respect the interests of all as free and equal. After all, insofar as liberal-egalitarian principles of justice just *are* those principles that are equally good for all, the presence of citizens protesting that their interests have been affronted gives us reason to rethink whether the principles selected in the first stage of justification really *are* in the universal interest. But in such a

⁴¹ Rawls, *A Theory of Justice*, p. 398.

⁴² Rawls is cognizant of the distinction between the reasonableness of principles and the likelihood of their realization from as early as 1951: “It should be noted that we are concerned here only with the existence of a reasonable method, and not with the problem of how to make it psychologically effective in the settling of disputes. How much allegiance the method is able to gain is irrelevant for our present purposes.” Rawls, “Outline of a Decision Procedure for Ethics,” in *Collected Papers*, p. 1.

⁴³ Rawls, “The Domain of the Political and Overlapping Consensus,” p. 473.

case, the objection to the conception would be that it was *mistaken*—i.e., a society regulated by it would be *unjust* on Stage 1 grounds, because out of sync with our considered convictions—*not* that the conception could not enduringly command free support. The latter would be only a *contingent consequence* of the conception’s moral unjustifiability, assuming citizens’ vigilant concern for their own interests, not the *source* of its moral unjustifiability.⁴⁴

One reason many have been drawn to the confirmation argument is that they misunderstand the point of political liberalism. George Klosko, too, subscribes to a version of the confirmation argument. He contends that the task of identifying plausible principles of justice should be accomplished by identifying the content of convictions that people already accept.⁴⁵ But in making this argument, Klosko misunderstands the point of political liberalism, supposing—falsely—that its central aspiration is to present principles that rest in straightforward concord with *existing* citizens’ belief sets. But political liberalism is not an anthropological hunt for existing agreement, nor is it an opinion poll. The idea of public justification in political liberalism appears in the theory *downstream* of the notion of seeing persons as free equals: the question is not whether a principle happens to be accepted by all, but rather whether a principle is acceptable to all *given the status of all as free equals*.

The confirmation argument might seem to be suggested by Rawls’s own discussion of “the strains of commitment” in *A Theory of Justice*. As Rawls writes, “[o]ne

⁴⁴ Quong makes this same criticism against Klosko (though not in the context of discussing stability) in a discussion over the distinction between the “internal conception” of political liberalism, which is not sensitive to the contingent beliefs of existing citizenries, and the “external conception,” which is. *Liberalism Without Perfection*, pp. 140-48. Gerald Gaus, as I read him, holds a sophisticated variation of the external conception, and accordingly would be sympathetic to a complex version of the confirmation argument; see *The Order of Public Reason*, pp. 232-332.

⁴⁵ George Klosko, *Democratic Procedures and Liberal Consensus* (New York: Oxford University Press, 2000), pp. 19-41.

cannot agree to a principle if there is a real possibility that it has any outcome that one will not be able to accept.”⁴⁶ The idea here is to impugn utilitarianism: it might seem well and good to seek the greatest good for the greatest number, but when circumstances arise such that the demands of this principle require the sacrifice of one’s fundamental interests for the sake of a minor increase in happiness incurred by the rest of the population, one will not “be able to accept” this principle. But it is not immediately clear what Rawls means by “able to accept.” On one interpretation—the interpretation that we must accept if we are to see the strains of commitment as part of the Stage 2 stability test—he is referring to facts about one’s psychology. So, we might imagine that one’s survival-instincts can be expected to kick in amongst those fated to die by some crude version of act-utilitarianism that demands the sacrifice of unpopular minorities, meaning that act-utilitarianism will not be able to endure freely. Those who expect such minorities to die quietly are in for a rude awakening. But while there is some plausibility to this predictive empirical claim,⁴⁷ it is not the best interpretation of the strains of commitment argument in Rawls. For, if utilitarianism really does require minorities to be sacrificed brutally in order to serve minor increases in happiness for everyone else, then the theory contravenes deeply held considered convictions. As such, the core vice of utilitarianism is not its Stage 2 instability, but its Stage 1 unjustifiability. Insofar as a conception must rest in concord with our considered convictions⁴⁸ in order to clear Stage 1 utilitarianism

⁴⁶ Rawls, *A Theory of Justice*, p. 159.

⁴⁷ For the idea that injustices are doomed to be short-lived *because they are injustices*, see Joshua Cohen, “The Moral Arc of the Universe,” in *The Moral Arc of the Universe and Other Essays* (Cambridge: Harvard University Press, 2010).

⁴⁸ Or, in the terms of the original position: secure the fundamental interests of the parties’ trustees.

should never have been “allowed” to progress to Stage 2. It never should have gotten out of the gate.⁴⁹

A different interpretation of the role of the strains of commitment argument *moralises* what it means to say that someone will not “be able to” accept a particular burden that the theory under consideration deems to be required of them. On this interpretation, to say that someone will not be able to submit to a particular burden is to say that she will not be able to submit to it in a way that is compatible with her self-respect, given her considered conviction that her life is of equal importance to the lives of others. But what is doing the work is not any brute fact of human psychology, but rather a normative judgement that any theory that requires people’s lives to be used principally as instruments of others’ pleasure defies a deeply held conviction. This makes Rawls’s choice of words less than ideal, but it explains how the strains of commitment could figure (as Rawls’s location of the discussion, in *Theory’s* Part I, suggests) as part of the reasoning of Stage 1.⁵⁰

Consider now a final version of the confirmation argument. Samuel Freeman argues that if principles of justice “defeat our primary purposes, prevent us from realizing important human goods, or require ways of acting that it is not consistently in our nature to perform, then surely this is relevant to the justification of a conception of justice.”⁵¹ So, if the *reasons* for predictable resistance to a conception of justice trace to

⁴⁹ Though his argument for this conclusion is somewhat different, the same conclusion—that utilitarianism is best understood as ruled out on *moral* grounds, not psychological grounds—is reached in Barry, *Justice as Impartiality*, pp. 61ff.

⁵⁰ My argument here presupposes that the central sense of stability is, indeed, the binary sense of stability (“stable or unstable”), not the gradated sense of stability (“more or less stable”). One reason to doubt that this is the right interpretation is because of Rawls’s discussion in which he contrasts the comparative stability of justice as fairness and utilitarianism (e.g., *A Theory of Justice*, p. 433ff). But it is important to see that this comparison appears in an argument about why justice as fairness is stable, not an argument about why it is *more* stable than utilitarianism.

⁵¹ Freeman, “Congruence and the Good of Justice,” p. 154.

any of these categories—if a conception’s demands are, in Freeman’s language, *self-abnegating*⁵²—this is good grounds to dismiss a conception of justice as implausible. The problem, however, is that *if* a conception of justice really did require us to live terrible lives, it would run afoul of our considered convictions—precisely a Stage 1 concern. Consider Plato’s proposal for children to be raised in community care.⁵³ It may seem like the fact that parents would naturally be inclined to resist such a policy is what constitutes a decisive reason against it. But we are bound to ask parents who would so resist: *why*? And the answer could not *alone* be, because it would happen to make them very sad. Rather, the answer, at its best, will be an account of the value of the parent-child relationship and the importance of securing the freedom and opportunities to pursue and enjoy such relationships.⁵⁴ Now, it is a vexed and important question, which I shall not attend to here, how empirical facts about human beings figure in the specification of their fundamental interests as free and equal citizens—the specification of interests, that is, used by liberal-egalitarians to formulate principles of justice.⁵⁵ Most likely different specifications will compete, as attached to distinct reasonable conception so justice, in the public political culture of liberal societies. But concerns of value are importantly distinct from the concerns of psychology and feasibility that Rawls seeks to understand, and should not be confused with them. That is our lesson from rejecting the confirmation argument.

⁵² Freeman, “Congruence and the Good of Justice, p. 155. Freeman is considering this objection to liberalism as traceable to Friedrich Nietzsche, *On the Genealogy of Morality*, translated by Carol Diethe (Cambridge: Cambridge University Press, 1994).

⁵³ Plato, *The Republic*, p. 136.

⁵⁴ E.g., Harry Brighouse and Adam Swift, “Legitimate Parental Partiality,” *Philosophy & Public Affairs* 37, 1 (2009): 43-80.

⁵⁵ For related discussion, see David Miller, “Political Philosophy for Earthlings” in *Political Theory: Methods and Approaches*, edited by David Leopold and Marc Stears (Oxford: Oxford University Press, 2008).

2.2. The purpose of moral norms

Consider a fresh argument as to why Rawls might have granted Stage 2 justificatory import. On this view, the point of the stability argument flows from the role Rawls assigns a conception of justice (or, for our purpose, a conception of a fundamentally just society) in constituting the “foundation charter”⁵⁶ or “public moral constitution”⁵⁷ of a political order.⁵⁸ Because the “social role” of principles of justice is to regulate human societies enduringly, we must finesse such principles’ demands to accommodate our nature, lest the principles not be able to fulfil their role.

Rawls appears to offer us grounds for this position in the following passage:

[C]onditions for justifying a conception of justice hold only when a basis is established for political reasoning and understanding within a public culture. The social role of a conception of justice is to enable all members of society to make mutually acceptable to one another their shared institutions and basic arrangements, by citing what are publicly recognized as sufficient reasons, as identified by that conception.⁵⁹

But Rawls here is referring to the fundamental *moral content* of the normative theory—the considered convictions that serve as the normative raw material at Stage 1. The concern with action-guidance thereby bears not on stability—which concerns the *acquisition* and *maintenance* of the sense of justice—but on what fundamental ideas are employed within the theory.

So we have reason to doubt this position’s exegetical plausibility. But more importantly, we have reason to doubt its philosophical plausibility. The idea that a public conception that rests in concord with our considered convictions will not “fulfil its role”

⁵⁶ Rawls, *A Theory of Justice*, p. 10.

⁵⁷ Rawls, “Kantian Constructivism in Moral Theory,” p. 305.

⁵⁸ Gerald Gaus, *The Order of Public Reason* (Cambridge: Cambridge University Press, 2011); Gaus, “On the Appropriate Mode of Justifying a Public Moral Constitution,” *The Harvard Review of Philosophy* (forthcoming); Cf. G.A. Cohen, *Rescuing Justice and Equality* (Cambridge, MA: Harvard University Press, 2008), pp. 274-343.

⁵⁹ Rawls, “Kantian Constructivism in Moral Theory,” *The Journal of Philosophy* 77, 9 (1980): 515-572, p. 517.

if people are unlikely to generate the conditions of its own freely willed support is an elaborate way of saying that people are *unlikely* to choose to live up to it. But the *unlikelihood* of a particular moral principle's realisation strikes us in ordinary morality as irrelevant to whether we *ought* to act in accordance with it. "What is it to hold a moral principle," asks Eric Beerbohm, "other than to have a durable commitment to *its* realization—not some paler version that is within arm's reach?"⁶⁰ Even if racist inclinations were naturally selected for as effective evolutionary means for group survival, it would be perverse to regard such facts as decisive reasons for abandoning our belief in the importance of integration and replacing it with a commitment to "separate but equal" social arrangements—or worse. And even if morally-directed pragmatic considerations of implementation spoke in favour of *second-best* "rules of regulation" less demanding than their moral counterparts (due to obstinate features of human psychology), we would regard this as a *compromise*, not an actual retraction of the initially identified moral standards, which stand. In such circumstances, there would be a shameful, publicly recognised gap between our publicly avowed moral constitution and our standing law. As G. A. Cohen argues, "What ought justice, or the principles of justice, to be?"⁶¹ is an incoherent question for those who believe in objective (or intersubjective) moral standards.⁶²

⁶⁰ Beerbohm, *In Our Name*, p. 163. This helps us see that stability is not plausibly some *additional value* to be *weighed against* the value of justice—such that we might be willing to sacrifice x amount of justice for y amount of stability (though we could easily imagine a pluralistic moral theory with that sort of character). And this could not be Rawls's own view, for he is concerned with the *stability of justice*—and a quite idiosyncratic understanding of what that involves—not stability *simpliciter* (using, for example, some definition drawn from political science).

⁶¹ Cohen, *Rescuing Justice and Equality*, p. 277.

⁶² And indeed we can regard it is an incoherent question while *still* accepting Rawls's claim (*pace* Cohen) that ideals of justice *are* (ideal) rules of regulation, in the sense of constituting the basic standards of interaction that ought to regulate our relationships as fellow citizens.

Perhaps, however, there is a more sophisticated version of this claim—that stability’s role relates to the practical function of moral principles—that merits our attention. Samuel Freeman offers a tantalisingly brief suggestion that the stability test serves an equivalent hurdle for the justification of a public conception of justice as the “ought implies can” test serves for the justification of single normative principle. Because the principle that “ought implies can” (OIC) is widely believed to invalidate normative requirements that do not meet its strictures, a stability argument grounded in this principle could explain how unstable conceptions could accordingly be invalidated on the basis of their instability. To evaluate this argument, we need to understand what OIC means.

OIC reflects the widely held conviction that if someone cannot Φ , it is not only false, but incoherent, to claim that he ought to Φ . To say that an agent is subject to a moral standard to Φ implies that Φ is within his choice set. And it implies this because of what it centrally *is* to say that an agent is subject to a moral principle: that he ought to *choose* to comport his conduct with that standard. Moral reasons are *practical reasons*: they are considerations that count in favour of agents choosing to act in certain ways instead of others. Thus moral principles bear on choices actually available to people.

This description serves to remind readers why OIC is intuitively plausible: if it were false, much of what we assume to be correct about the activity of moral deliberation and criticism would unravel. It makes no sense to claim that agents are subject to standards that they could not possibly live up to. I shall not defend the principle’s plausibility any further. What I shall do is inquire what it means to say that a certain choice is unavailable to an agent. Two kinds of normative principles are uncontroversially invalidated by OIC: *logically impossible* principles and *physically impossible*

principles. No single, justified normative directive can require me to draw a circle whose circumference is not that circle's diameter multiplied by π , nor can one require me to stop a falling object telekinetically. But clearly *these* are not the concerns that Rawls has in mind when he writes of a conception's tendency to generate its own freely cultivated and freely enduring support, given human nature, so some other sense of "can" in "ought implies can" must be at stake if Freeman's suggestion is to be vindicated.⁶³

One promising further candidate—one that connects up to the sorts of concerns that Rawls's treatment of stability has in mind—concerns *psychological* impossibility. What does it mean for something to be psychologically impossible? One thing it certainly does *not* mean is "psychologically difficult." However unlikely it is that some people will choose to vote for policies that increase the foreign aid budget, it does not follow that they *cannot* choose to vote this way. It simply means that they probably *won't*. So what does psychological impossibility really involve? There is a large and complex literature that attempts to explicate the answer to that question, and I could not hope to do justice to it here.⁶⁴ Nor do I need to, for my concern is not with psychological impossibility *simpliciter*, but with the proposition that a certain *proposed moral demand* is psychologically impossible.

How should we think about the very *idea* that a moral demand is psychologically impossible? Recall that there are two functions of moral theory: the *deliberative* function,

⁶³ The discussion here focuses on the matter of what Derek Parfit would call *deep* impossibility, not *technical* impossibility. See Parfit, *Reasons and Persons* (Oxford: Oxford University Press, 1984), p. 388.

⁶⁴ See, for example, Stuart M. Brown, Jr., "Does Ought Imply Can?" *Ethics* 60, 4 (1950): 275-84; D. G. Collingridge, "'Ought-Implies-Can' and Hume's Rule," *Philosophy* 52, 201 (1977): 348-51; Frances Howard-Snyder, "'Cannot' Implies 'Not Ought'," *Philosophical Studies* 130, 2 (2006): 233-46; Robert Kane, "Deontic Acts, Frankfurt-Style Examples, and 'Ought' Implies 'Can,'" *The Journal of Ethics* 4, 4 (2000): 357-60; John Kekes, "'Ought Implies Can' and Two Kinds of Morality," *The Philosophical Quarterly* 34, 137 (1984): 459-67; Walter Sinnott-Armstrong, "'Ought' Conversationally Implies 'Can,'" *The Philosophical Review* 93, 2 (1984): 249-261; Robert Stern, "Does 'Ought' Imply 'Can'? And Did Kant Think It Does?" *Utilitas* 16, 1 (2004): 42-61.

which helps agents decide what to do; and the *critical* or *evaluative* function, which helps agents decide whom to praise or blame for certain courses of action. This dissertation is more or less entirely restricted to principles' deliberative function, as it takes up the first-personal perspective of reasonable citizens committed to liberal justice. Remembering that helps us pose the question I have in mind more instructively: what kind of a claim am I making when I say, "Yes, I agree that this moral standard is vindicated by the best arguments; I agree that a failure to comply with it would affront the status of others as free and equal. But I just cannot live up to it; it is psychologically impossible for me to choose to do it"? Without wishing to deny that there are cases of genuine pathology, I regard this statement as generally incoherent. To see something as morally required just *is* to see oneself as properly moved to do it. Such a statement is incoherent for the powerful reason that *if* an agent regards herself as subject to a moral requirement, she *presupposes* that she is capable of living up to it. And the mere fact that many human beings would be *unlikely* to live up to it—even if that unlikelihood were traceable to some common element of their biology or psychology—does not make a shred of difference to the determination of capability.⁶⁵

This famous argument traces, most influentially, to Kant. Indeed, Robert Stern has persuasively argued that Kant never intended his dictum of OIC to be marshalled for the purposes of constraining the demands of morality when it comes to matters of

⁶⁵ My argument here is compatible with much of what is said in David Estlund, "Human Nature and the Limits (if any) of Political Philosophy," *Philosophy & Public Affairs*, 39, 3 (2011): 207-237. Our arguments share the fact that we think discoveries about human beings' common traits cannot alter the content of our convictions. However, Estlund never advances the argument—which I advance here—that part of what it is to see oneself as an agent is to see oneself as *capable* of living up to morality's demands. Estlund's account implies, perhaps reflecting compatibilist commitments on his part, that people may be subject to moral requirements even if they cannot live up to them. But my account—centred on the first-personal, deliberative point of view—finds that position incoherent and thus an implausible basis for the argument I advance here.

psychology; he meant the dictum *the other way around*: if I *ought* to do it, it must be true that I *can* do it.⁶⁶ As Kant writes:

Impulses of nature, accordingly, involve obstacles within the human being's mind to his fulfillment of duty and (sometimes powerful) forces opposing it, which he must judge that he is capable of resisting and conquering by reason not at some time in the future but at once (the moment he thinks of duty): he must judge that he can do what the law tells him unconditionally that he ought to do.⁶⁷

Kant grounds his argument in his controversial transcendental metaphysics. But we can account for it straightforwardly through reflection on the phenomenology of moral experience. It is generally an inescapable feature of our first-personal engagement with the world in liberal democracies, whatever else we believe about matters of deep truth and determinism, that we take ourselves to be in control of our psychological capacities, at least insofar as they bear on the performance of morally required actions. Even compatibilists, who explicitly accept that an ordinary person may well be psychologically incapable of living up to some ordinary moral demand, would concur that *from her first-personal perspective*, she nevertheless presupposes herself to be capable of choosing whether to live up to the moral demand or not. And that is enough, at the level of morality's *deliberative* function, to conclude: the idea of psychological impossibility (or, at least, *motivational* impossibility) is *deliberatively irrelevant* when it comes to an agent's assessment of whether she ought to live up to a moral principle that she believes is required of her. Now, simply because she *can* live up to morality does not mean she *will*. She presupposes herself to have the freedom to choose wrongly. And her phenomenology is not one of existing in a noumenal vacuum; she experiences temptations that tempt her to act wrongly. But if she succumbs to them—and this is the crucial point—she will understand herself to have *chosen* to succumb to them, to have *permitted* herself to

⁶⁶ Stern, "Does 'Ought' Imply 'Can'? And Did Kant Think It Does?" pp. 52-61.

⁶⁷ Immanuel Kant, "The Metaphysics of Morals," p. 513.

acquiesce. To be sure, different comprehensive doctrines will ground this phenomenology differently. But it is so woven into our moral experience—our sense of ourselves as endowed with the first moral power—that any doctrine that *denies* this account of our phenomenology (or one relevantly similar) has an implausible understanding of how we experience our sense of justice.

We can finally return to the issue of stability. Our initial question was whether we could identify any basis in the *function* of moral norms—their role in guiding action—to vindicate stability’s justificatory role. From the perspective of an agent who is committed to a particular conviction, should it matter to the judgement of whether that conviction *binds me* whether I or others are *likely* to comply with its dictates freely, from our sense of justice without incentive? It should not. This applies, importantly, regardless of whether we interpret Rawls’s stability test as one about the likelihood of justice *or* the possibility of justice; both are immaterial.

I have implicitly been addressing only one element of stability as Rawls understands it. There are, recall, four elements to stability: the issue of whether members of a well-ordered liberal democracy will *acquire* a sense of justice; the issue of whether they will come to affirm *liberal convictions* to which their sense of justice will be indexed; the issue of whether they will grant their sense of justice *regulative primacy* over their practical reasoning; and the issue of whether they will maintain reasonable beliefs and continue to apply them intelligently to fresh issues and concerns. I have here focused strictly on the third issue, but I believe my argument applies, *mutatis mutandis*, to the other elements, as well. To be clear, the issue is whether we, here and now, should regard it as possible for members of a well-ordered liberal democracy to do these things. But how could those of us, here and now, who affirm liberal convictions *deny* that it is possible to acquire a sense of justice and to be educated in accordance with liberal convictions? For,

living liberals are living proof—liberals that *haven't* been educated in a well-ordered society. That we are capable doing these things does not, of course, mean that we *will*—for we are fallible. As for sustaining liberal convictions, we need only consider the testimony of people who abandoned previously illiberal convictions—who look back in horror at their past selves, outraged that they treated people so badly—to see that it, too, belongs to the phenomenology of agency to see oneself as capable of *understanding* what it takes to value others.

2.3. The argument from hope

The third argument I shall consider is one recently advanced by Paul Weithman. Weithman's analysis is meticulous and sophisticated, and I cannot hope to do justice to its subtlety here. But its main contours are as follows:

- (1) The point of the stability argument is to shore up our faith in human nature—a faith without which we will not be able to sustain our motivation to pursue justice from *our* current vantage point in unjust societies.
- (2) If we could imagine how a just social order might reproduce itself freely, given facts of human nature as they are, this would be sufficient to give us faith.
- (3) A social order could reproduce itself in this way, given human nature, only if its members each believed that acting justly was prudentially rational; only then would they assign their sense of justice regulative primacy in their practical reasoning.
- (4) Members will believe that assigning their sense of justice regulative primacy is prudentially rational only if they have assurance that *others* will do the same.⁶⁸

⁶⁸ Weithman, *Why Political Liberalism*, pp. 327-35.

- (5) Members will have that confidence primarily because they will see that all members of the social order embrace a comprehensive doctrine that itself affirms justice as part of their good, and they will see others continually reaffirming their commitment to justice through the use of public reason.

This argument is fascinating. But it does not help us to vindicate the assignment of *justificatory* significance to stability's achievement. We can see so by clarifying (3), tracing its implications for (4), and then concluding by examining (1).

Start with (3). It might seem *guaranteed* that members of a well-ordered society would regard being just as part of their good “[B]eing the sorts of persons they are,” Rawls writes, “the members of a well-ordered society desire more than anything to act justly”.⁶⁹ Now it would, Rawls thinks, be “trivial” to ask whether persons so described would find being just consonant with their good.⁷⁰ The real question for stability is whether those persons would still regard being just as consonant with their good when *ignoring* their desire to be just for its own sake—when thinking within “the thin theory of good” (or, in the terms of *Political Liberalism*, the content of their comprehensive doctrine *minus* its intrinsic valuation of the sense of justice). If they would *not*, they would face temptations to become the sort of people who choose whether to act justly case-by-case. If the citizens of a well-ordered society all resolved accordingly, the justice of their relations would soon unravel.

On to (4): why does *assurance* that others will choose to assign primacy to their sense of justice make the relevant difference to what citizens resolve within the thin theory? The answer can be gleaned by examining a passage in Rawls:

⁶⁹ Rawls, *A Theory of Justice*, p. 498.

⁷⁰ *Ibid.*, p. 498. For criticism of Rawls on a related point, see Brian Barry, “John Rawls and the Search for Stability,” pp. 884-85.

For although men know that they share a common sense of justice and that each wants to adhere to the existing arrangements, they may nevertheless lack full confidence in one another. They may suspect that some are not doing their part, and so they may be tempted not to do theirs.⁷¹

There are two issues that arise here. The first is that citizens are worried about being played for a “sucker,”⁷² paying taxes and abstaining from self-serving behaviour while others free ride. The second is that citizens are worried that continuing to be just will be *futile*; if others are unlikely to continue to do their part, the goods provided by justice—the goods that make a just world desirable even within the thin theory—will not be realised.

What I want to emphasise is that the inclinations about which we are worried in each of these cases are *morally pernicious* inclinations. Take the first worry. Even if our relations are impaired with some of our citizens when they choose to free ride, pursuing justice is *morally required* under the natural duty of justice, not morally optional. Fundamental duties of justice are not duties of fair play. Simply because we are envious of others’ unjust enrichment and are properly angry with them for failing to do their part, this does not give us the right to *exacerbate* the defunding of schools and hospitals that results from, for example, cheating on taxes. Secondly, consider the inclination to *dispense* with the maintenance of a presently well-ordered society out of a concern with futility. Now, I would argue that assurance is an existence condition of institutional duties of justice since citizens are not morally required to fund institutions if they have no idea whether those institutions will do their job. But note that even without assurance, justice simply requires pursuing the just objective non-institutionally (so, without good grounds for thinking that the tax office is actually going to get one’s tax dollars to the homeless

⁷¹ Rawls, *A Theory of Justice*, p. 211. I return to this passage in Chapter Five, since it relates to Rawls’s limited comments on punishment.

⁷² G. A. Cohen, *Rescuing Justice and Equality*, p. 351.

shelters, one ought simply to bring the cash directly to the shelters⁷³). At any rate, worries about lack of confidence in institutional effectiveness can be allayed if the state compiles and distributes democratically audited “public compliance records”—indicating the statistics on compliance—to give people the confidence they need. Those who would *still* defect from being just in such circumstances would, I submit, be acting wrongly.

In light of what I argued earlier—about what it *is* to see oneself as subject to a moral conviction, and what it *is* to regard oneself as an agent—it is difficult to see why, within the first-personal perspective, we would hold the *justification* of our deepest convictions hostage to the empirical likelihood that these pernicious inclinations will, as Rawls hopes to show, “take care of themselves” in a well-ordered society. Even without considering Weithman’s broader argument, its reliance on an aspiration to accommodate the pernicious inclinations I have just described gives us reason, in light of our analysis thus far, to doubt its success in rendering stability a justificatory condition of a public conception.

In any case, let us go on to consider Weithman’s first premise, on the subject of hope. Weithman’s concern here is partly religious. If we cannot show that humans are naturally good—naturally likely to respect one another’s rights out of genuine concern for one another’s value, not out of fear or deception—we will not be able to make sense of the *Book of Genesis* 1:31: “God saw everything that he had made and, behold, it was very good.”⁷⁴ But the fundamental concern is more ecumenical, concerning the preconditions of human motivation. The basic insight is this: if we have good grounds for thinking that human beings are hardwired to hate, oppress, and kill each other—and

⁷³ For support for this line of argument, see Aaron James, “Mutual Assurance Problems in Global Context,” in *Hobbes Today: Insights for the 21st Century*, ed. Sharon Lloyd (New York: Cambridge University Press, forthcoming.)

⁷⁴ Quoted in Weithman, *Why Political Liberalism*, p. 367.

human history seems to provide precisely such grounds—we will not be able to sustain our commitment to justice. It is worth quoting Kant at length here:

Is the human race as a whole to be loved, or is it an object such that one must regard it with vexation, for which one indeed wishes everything good...but of which one must never expect this, so that one must prefer to avert one's eyes from it? The reply to this question rests upon the answer one gives to another: Are there in human nature predispositions from which one can gather that the race will always progress toward what is better and that the evil of present and past times will disappear in the good of future times? For in that case we could still love the race, at least in its constant approach to the good; otherwise we should have to hate or despise it...For, however, one may try to exact love from oneself, one cannot avoid hating what is and remains evil, especially in deliberate mutual violation of the most sacred human rights...⁷⁵

Rawls offers a gloss on this discussion in his *Lectures on the History of Moral Philosophy*, discussing Kant: “[Kant] believes that we cannot sustain our devotion to the moral law...unless we firmly believe that its object is possible.”⁷⁶ As Weithman points out, Rawls is not referring to the members of the well-ordered liberal society in this passage. He is referring to *us*, here and now. The insight is that *if* we are not able even to *imagine* how a well-ordered society might reproduce itself freely over time, then the most important ingredient of a commitment to justice as it operates in gravely unjust societies will be destroyed. That ingredient is *hope*: hope for a more just future.

In the Introduction to this dissertation, I quoted Kant's worry that our sense of ourselves as capable of justice—the very sense I elaborated above when discussing “ought implies can”—might be illusory: that we might *think* we are capable of justice, for we can see the force of moral reasons, but we might simply not turn out to be the sorts of creatures who can regulate their conduct by those moral reasons with any reliability. In such a case, human beings would turn out to be like lions or wolves—only more

⁷⁵ Kant, “On the Common Saying: That may be correct in theory, but it is of no use in practice,” pp. 305-306. Weithman does not flag this particular passage in Kant, but it appears to be the source he and Rawls have in mind.

⁷⁶ Rawls, *Lectures on the History of Moral Philosophy*, pp. 10ff.

dangerous, and equipped with a tragic capacity to recognise that the mayhem they unleash upon others' interests is wrongful. This is precisely the fear that is under discussion now. Kant thinks we can allay this worry by offering a speculative empirical vision of how human nature might be congenial to justice.⁷⁷ This is a strategy he derives, I suspect, from Jean-Jacques Rousseau, whose hypothetical history traces human vice not to our nature, but to distortions in our self-understanding generated by defectively arranged institutions.⁷⁸ Rawls's stability argument thus follows in this tradition: by helping us see how a well-ordered society might sustain itself, our "reasonable faith" in the compatibility of enduring justice with human nature is secured.⁷⁹

Now I do not wish to downplay the importance of hope in human motivation. If we know that we are liable to become morally paralysed by the worry that human beings are unlikely to will justice freely, we should indeed strain to find these sorts of arguments if they will help to allay our depression. All this is true. But it does not touch the *justification*—the morally binding character—of our moral aspirations. It fails to account for stability's justificatory significance.

Firstly, the idea that we need hope to sustain our commitment to justice presupposes *that we have already identified the relevant standard of what justice is*. Only if we *already* know what justice is would it strike us as tragic if we believe human beings would never choose to live up to it. Joshua Cohen registers an intense belief in the motivational significance of hope throughout his writings, and yet he nevertheless says that the question of hope is the question: "*given the demands of morality* and what we know about

⁷⁷ Kant, "On the Common Saying," p. 307.

⁷⁸ Rousseau, "Discourse on the Origin of Inequality," 60-81. Here I follow Joshua Cohen, *Rousseau: A Free Community of Equals* (Oxford: Oxford University Press, 2010), pp. 129-130.

⁷⁹ Rawls, *Political Liberalism*, p. 172.

how the world does and might work, what sort of world, we ask, is it reasonable to hope for, and to strive to achieve?”⁸⁰

Secondly, however motivationally important hope might be, we have reason to doubt that a public conception’s justification could turn on something as *vague* as whether it could reasonably be hoped for. What exactly does it *mean* to see human nature as “not unfriendly”⁸¹ toward justice? Even if there is such a thing as a homogenous human nature, it is plausible to think that some elements of it are congenial to justice, and some are not. Our naturally selected tendency toward reciprocity is an example of the former, and our naturally selected tendency to be fearful of those different from us is decidedly an example of the latter. Are we to search through the list of attributes conferred upon us by evolution, identify the amounts of psychological pressures they adduce, and see whether friendliness or animosity toward justice results from the balance of those pressures?

Thirdly, vagueness aside, it defies our sense of ourselves as free to hinge our moral aspirations on such empirical self-assessments. Suppose we learn that we are contested souls, with a host of forces within us disposing us toward justice, and an equally balanced host of forces within us disposing us toward injustice. The right response—as I shall soon argue—is to galvanise and fortify the former, not to suspend one’s commitment to justice out of fear that we are by nature hostile toward it. Indeed, even if our motivational structures were such that just orders would likely to be sustained only for a hundred years at a time before decaying, why should we give up those hundred years? Those whose lives stand to be blighted by injustice would have a reasonable objection to our apathy. “Why aren’t you working hard to save us from genocide?”

⁸⁰ Joshua Cohen, “Minimalism About Human Rights: The Most We Can Hope For?” *The Journal of Political Philosophy* 12, 2 (2004): 190-213, p. 191. Emphasis added.

⁸¹ Weithman, *Why Political Liberalism*, p. 364.

“Well, you see, given facts of human nature, it is unlikely that the achievement of justice would endure...” Given that we have *some* evidence that we might be creatures capable of justice—the evidence provided both by limited history and our normative self-understanding as autonomous—it is implausible for us not even to *try*. Adapting Dante: there is a special place in Hell for entities capable of giving justice to each other who cynically and self-servingly refuse.⁸²

§3. Stability Reborn

I have argued that stability—the potential of a conception of justice to generate its own freely willed support and maintenance—is justificatorily irrelevant. Our fundamental moral convictions do not retreat in response to empirical discoveries that our nature could be hostile to them; we neither dilute nor abandon them in response to the judgement that our preferred public philosophy—in our case, the ideal of a well-ordered society—is likely to face resistance. Thus the normative ideal sketched in Chapter One remains intact. Our normative self-understanding corroborates our status as capable of living up to it. Many political philosophers will be disposed to agree with this conclusion,

⁸² This saying is commonly attributed to Dante Alighieri, “The Inferno,” in *The Divine Comedy*, translated by Henry Wadsworth Longfellow (Leipzig: Bernhard Tauchnitz, 1867), Canto 3, lines 35-42, p. 20, referring to a condemnation of those indifferent between good and evil:

And I, who had my head with horror bound,
Said: “Master, what is this which now I hear?
What folk is this, which seems by pain so vanquished?”
And he to me: “This miserable mode
Maintain the melancholy souls of those
Who lived withouten infamy or praise...”

It has been popularised by Madeleine Albright, who said a variant of it at a keynote speech for the WNBA in 2006.

and accordingly resolve to stop thinking about the question of stability altogether.⁸³ That, however, would be a profound mistake, and for two reasons: one concerned with the moral importance that justice be *enduring*, another concerned with the moral importance that justice be *freely willed*.

The arguments I have offered have established that we must see ourselves as capable of living up to justice's demands. But a vital question remains: *will* we? There are two versions of this question. Firstly, will we choose to *transition* from our presently unjust circumstances to just circumstances? Secondly—imagining ahead to a possible future that realises our ideal of a well-ordered society—would members of a well-ordered liberal democracy choose to *sustain* their basically just relations? With admitted artificiality, I want to focus attention directly on the second question, since it avoids certain complications while simultaneously providing us with qualified guidance as we think through the first question. So let us take up the first-personal perspective of citizens in a hypothetical well-ordered liberal democracy. Having read their Rawls, they are aware of his hope that such a society would be able to *sustain* itself over time. But they are not so sure. “How could Rawls have been so confident,” they wonder, “in light of his epistemic vantage point in an unjust society, that he could distinguish *contingent threats* to the successful operation of our first moral power—those generated by unjust institutions—from *enduring threats* to its successful operation? How would *he* know what we would face?” Such citizens have seen the horrors of human history; they have

⁸³ For example, see Jürgen Habermas, *The Inclusion of the Other*, pp. 90ff; Brian Barry, “John Rawls and the Search for Stability,” pp. 901-902; and Adam Swift and Stephen Mulhall, *Liberals and Communitarians*, p. 240ff. But their endorsements of the disconnection between justification and stability (which often focus on narrow concerns about the overlapping consensus in particular, rather than the general problem) do not explore with sufficient care the different possible grounds for *resisting* such a disconnection, as I have pursued in this chapter. Given how ready many political philosophers are to take Rawls's views on a matter as *prima facie* correct, careful scrutiny why he was wrong about such a central theme of his work has considerable payoffs.

reviewed the psychological research on human fallibility; and they have experienced making mistakes about moral reasoning and succumbing to temptations to act wrongly. To be sure, their society is not riven with envy and greed to the same extent as societies of the past; the intensity of the inegalitarian impulse has waned, just as Rousseau predicted. But such impulses nevertheless remain.⁸⁴ They therefore worry that even though they see themselves as *capable* of living up to the ideal of freely willed justice—*ex hypothesi*, they *have* been living up to it—they nevertheless *will not* continue to so. And they are not content to cross their fingers and simply *hope* that they will be able to sustain their sense of justice. They want to plan for it. They seek to take *proactive measures* to render justice *enduring*. This is the first crucial way in which stability retains its significance: those who have a commitment to justice necessarily care about whether justice will endure, and are prepared to work hard to see that it does. And so they ask their political philosophers: *what should we do?*

Consider a proposal, which I will call *DRONES*; it is outlined by Warren Quinn:

Let us imagine ourselves existing at some time in the future when...[b]eing scientifically very advanced...we are together capable of making fantastically complex devices that can (at least as well as we can) detect wrongdoing in our new community, identify and apprehend those who are responsible, establish their guilt, and subject them to incarcerations (and perhaps other evils) that I shall call mechanical punishments, or m-punishments for short. Let us further imagine that we have lost whatever taste we once had for retribution and are interested only in protection. The devices attract us, therefore, in their deterrent rather than in their retributive capacities.⁸⁵

⁸⁴ Aaron Ben-Ze'ev, "Envy and Jealousy," *Canadian Journal of Philosophy* 20 (1990): 487–516, and "Envy and Inequality," *The Journal of Philosophy* 9 (1992): 551–581. For discussion of Ben-Ze'ev, see Patrick Tomlin, "Envy, Facts, and Justice: A Critique of the Treatment of Envy in Justice as Fairness," *Res Publica* 14, 2 (2008): 101–16.

⁸⁵ Quinn, "The Right to Threaten and the Right to Punish," p. 337. Quinn credits the idea of such a device to Alexander, Lawrence, "The Domsday Machine," *The Monist*, 63, 2 (1980): 199–227, and James Buchanan, *The Limits of Liberty* (Chicago: University of Chicago Press, 1975), p. 95.

The members of the well-ordered society take up *DRONES* for debate. It strikes some citizens as precisely the right course of action. “A commitment to justice,” they say, “involves doing *whatever it takes* to ensure that the right distribution of freedom and resources is achieved—to make the state of affairs definitive of ‘justice’ real—even if it involves deploying robots to buzz through the skies, terrorising the population into submission.” Having been convinced that *freely willed* stability has no justificatory significance, others nod their heads. But some dissenting citizens remind them: “But wait! Remember what norms of justice *appraise*: not mere distributions, but the moral quality of the human relationships that underlie them.”

Albeit *DRONES* is a crude proposal that skates over subtleties that will preoccupy us in due course, the dissenters’ objection helps us see the second vital way in which stability remains normatively significant. In Chapter One, I declared my allegiance to the widely believed normative view that justice is fundamentally a matter of human relationships rather than distributions, as well as the exegetical view—held by many of Rawls’s students—that Rawls himself embraced this relational understanding of what justice involves. If that is right, it helps us see why the *freely willed* maintenance of certain rights-respecting, opportunity-enabling institutions matters so much: because part of what it *is* for people to be in a relationship with others is for them to undertake it *for the right reasons*. As T. M. Scanlon writes, to conform one’s conduct freely to mutually acceptable principles “puts one in a relation of mutual recognition with others, a kind of relationship that is worthwhile for its own sake as a way of living with others.”⁸⁶ But this relationship cannot obtain if it is only maintained by threats of coercion; it ceases to be the right kind of relationship.

⁸⁶ Scanlon, *What We Owe to Each Other*, p. 162.

I believe that this is the best understanding of why the freely willed achievement of aspirationally just institutions matters so much: because only when we *freely affirm* one another's status as free and equal is the value of justice fully realised. This claim will do some work at various places down the road in future chapters, so I shall pause on it momentarily. Note some features of what valuable relationships consist in. Firstly, a valuable relationship is what Niko Kolodny calls "attitude-dependent."⁸⁷ Part of what gives the relationship the value that it has is that we affirm it as one worth being in. We know this to be plainly true from personal relationships with friends and lovers; we experience a saddening alienation when we learn that they do not wish to be in the relationship, or do not want to regulate the relationship by the standards that make it the kind of relationship that it is, or that we want it to be. Importantly, as the case of our relationships with family members makes clear, such affirmation does not depend on the relationship's status as voluntarily chosen. Nor does the importance of free affirmation presuppose that the relationship is morally optional, as the relationships that we ought to have with fellow citizens decidedly are *not*.⁸⁸

Secondly, a valuable relationship is one in which the conduct of the parties to the relationship is regulated by the relevant standards. In the case of just civic relationships, the relevant standards are those given by the relationship's purpose: to generate a social world in which people can develop and exercise their two moral powers. If our civic relationships become regulated by different standards—for example, those suitable to the perpetuation of class divisions or racial domination—it ceases to be a just relationship,

⁸⁷ Niko Kolodny, "Love as a Valuing Relationship," *The Philosophical Review* 112, 2 (2003): 135-89, p. 149.

⁸⁸ Here I follow Andrew Lister, "Reciprocity, Relationships, and Distributive Justice," *Social Theory and Practice* 39, 1 (2013): 70-94, p. 87; T. M. Scanlon's discussion of relationships in *Moral Dimensions: Permissibility, Meaning, Blame* (Cambridge: Harvard University Press, 2008), p. 142. The ideal of just civic relationships is an ideal toward which we are required to strive.

just as the relationship between two old friends ceases to be a friendship when one of them decides to start being cruel to the other.

Thirdly, a valuable relationship is one in which the relevant parties have reasonable grounds for believing that the relationship has the other valuable properties I have mentioned—they know, that is, that it is the *kind* of relationship that they aspire for it to be. It is not enough for the right conduct to obtain. Consider three examples. Suppose a father overhears his children whispering about how eager they are for him to die so that they can collect his money. Even if their relationship carries on as usual, regulated by the same standards of conduct, the father's discovery of his children's attitudes has illuminated to him that it was not the sort of relationship he thought it was. Suppose my dear friend, Chris, happens to acquire a bizarre illness according to which he will die unless he remains my friend. In such a case, it becomes very difficult to know whether he is remaining my friend out of his love for me, or because of his fear of death. Consider a wealthy woman, Rita, who prefers only to date wealthy men since only then can she believe that they are in the relationship for the right reasons. These examples testify to the importance of knowing what others' attitudes are. Something similar applies, *mutatis mutandis*, to our civic life. If we believe that the only reason our fellow citizens are not putting us in concentration camps is because they are afraid of punishment, our relationships do not secure their full value. This raises, however, a crucial problem: how do I know that the people with whom I am standing in a relationship have the relevant attitudes toward me? How do I know, to use our case, that my relationship with my fellow citizens really is a *just* civic relationship? The answer: if we place people in conditions in which their conduct is a free expression of their attitudes,

we will have excellent grounds for knowing whether they affirm us or not.⁸⁹ This is not to say that there are never occasions in which we are not willing to *risk* finding out what their real attitudes are; but my point is that those occasions represent a justified compromise of our *ideal*, according to which we aspire to honour freely one another's moral status. This is one of many grains of truth in the idea that "justice must be seen to be done."⁹⁰ We must be able to *see* one another as moved by the right reasons.⁹¹

This exposes the trouble with *DRONES*. Its undesirable character does not flow from the controversial thesis that rightful action only has value when conducted in the teeth of countervailing motivation. Nor does it flow *just* from the idea that morally rightful action only has intrinsic value when freely performed—a fact which is true, but perfectly compatible with *DRONES*, since many people could presumably still choose for the right reasons. What's wrong with *DRONES* is that, in the world it describes, we would have excellent grounds for *doubting* that others are engaged in the relationships for the right reasons. A vulnerable minority who has good reason to wonder whether *DRONES* is what is keeping them safe from oppression experiences an impaired form of relation with their fully citizens.

I have outlined the two ways that stability retains its normative significance, despite its justificatory insignificance. But the two ways seem to rub against each other.

⁸⁹ This argument can be advanced in many different forms. For example, for those who embrace the conception of *justice as fairness*, this argument can be cast as one about the existence conditions for the social bases of self-respect.

⁹⁰ I shall not go into the details of what this implies for issues of personal choice, but see Andrew Williams, "Incentives, Inequality, and Publicity," *Philosophy & Public Affairs* 27, 3 (1998), pp. 225-247. Note that this thesis does not imply that *every* morally required action must satisfy this sort of publicity test. What I say concerns our understanding of the *general character* of the relationship, not of every one-off action. Hence this argument is not incompatible with G. A. Cohen's rejoinder to Williams in *Rescuing Justice and Equality*, pp. 344-71. I cannot undertake a full defence of the nature of publicity here. For such a defence, see Thomas Christiano, *The Constitution of Equality: Democratic Authority and Its Limits* (New York: Oxford University Press, 2008), pp. 51ff.

⁹¹ The account follows closely T. M. Scanlon's discussion of relationships in *Moral Dimensions*, especially pp. 123-41.

On the one hand, we are committed to working hard to make sure that others' rights are respected in the face of our fallibility. On the other hand, we are committed to making sure that our relationships live up to their ideal form. These two commitments collapse to present a single challenge: *what should we do to increase the likelihood that our first moral power will operate effectively, so that we can realise the social world that is our ideal?* I stand for the view that normative political theory should fasten its attention on this question. For, only when we have articulated its answer from the perspective of the citizens of our ideal well-ordered society have we completed our specification of that ideal.

Importantly, this question is not a question we ask when thinking about stability in the way Rawls invites us to. After all, (one version of) the claim that a conception of justice is stable *just is* the claim that those living in a society regulated by it will naturally acquire a sense of justice that forms an enduringly regulative part of their psychology. On this view, if citizens of a well-ordered society even need to be *asking* themselves what they should do to shore up the first moral power, this is proof that this conception is unstable. Within the first-personal perspective of agents committed to justice, however, this *manner* of conceiving stability—as some sort of sociological exercise, external to the thoughtful operations of citizenship in the a liberal democracy—is implausible. Indeed, the seed of its implausibility is already in Rawls. Earlier in this chapter I examined some different interpretations of Rawls's political turn. We rejected the Public Basis View as an exhaustive explanation of Rawls's political turn, since it does not concern questions of stability. We also rejected the Empirical Prediction View. Instead we embraced the view that his political turn was motivated by an appreciation of the central place of reasonable pluralism within a free society. But there is a *fourth* possibility—one that retains some of the distinctive features of the Public Basis View. On this Revised Public Basis View, the stability argument Rawls delivers in *A Theory of Justice* is defective not simply because it

defies the kind of world that is certain to result from liberal institutions (one with reasonable pluralism); it is defective because we, together as citizens, regard it as *a morally objectionable* understanding of how the stability of our relations is to be secured. It is defective because it disrespects us as free and equal. On this view, the question of stability is *itself* conceived not as some sociological exercise external to the operations of thoughtful citizenship, but as *part* of the questions that we, together, must answer first-personally as democratic citizens. It is citizens' shared *plan* of how they see their commitment to justice as enduringly possible given the sorts of beings they are. Call this shared plan *citizens' stability charter*. On this view, the overlapping consensus's role as a stabilising force is not a mere *prediction*. It is part of our *strategy* for how we are going to make justice an enduring achievement: by continually nourishing the roots of our support for justice by keeping our comprehensive-doctrinal communities alert and committed to questions of public right, thereby galvanising and emboldening our first moral power and reminding ourselves that it is part of our good.

Note how this approach to stability *breaks free* of the constricting dichotomy between the imposed stability of *DRONES*—secured at the barrel of a gun through institutions specifically designed for preventing injustice by coercing and manipulating citizens—and what Rawls termed stability traceable only to “forces within the system”—secured organically through the spontaneous development and natural exercise of citizens' sense of justice, with no specially designed institutions necessary.⁹² For, there is a promising middle ground: *stability secured with the assistance of specially designed civic practices that have the aim of fortifying citizens' sense of justice.*⁹³

⁹² See Section 1.1.

⁹³ Consider Rawls's passage: “Perhaps we think there are two separate tasks: one is to work out a political conception that seems sound, or reasonable, at least to us; the other is to find ways to bring others who reject it also to share it...if need be prompted by penalties enforced by state

This is the basis of a radically different approach to the matter of stability. Rather than serving as a justificatory criterion set by philosophers on our convictions' justification, freely willed stability is a *practical challenge* set by our convictions *for us*. By answering this challenge, we acknowledge our fallibility—we take seriously the prospect that even those living in a well-ordered society will face challenges from their nature—but we do not respond by giving up on the ideal of just social relationships with one another, insisting instead that we must settle for a social world that endures only through force. That is not to deny that force is never appropriate, and it is not to deny that there are non-ideal circumstances in which coercion must play a significantly larger role than it would in our vision of a well-ordered society. What it *does* deny is that sober acknowledgment of our own fallibility automatically forces us to dilute or abandon our ideal of just social relationships. For, there is a fresh possibility: *reckoning* with our fallibility.

The stability charter of a well-ordered society is by its nature a large and diverse collection of norms, strategies, and mechanisms—both official and informal. Different elements of the charter are properly conceived and justified in different ways. In the remainder of this chapter, following what David Estlund has called the “the method of provisional leap,”⁹⁴ my aim is simply to sketch rudiments of an organising framework for thinking about how to conceive and justify a subset of elements of the stability charter I shall call *fortificational duties*: duties that citizens have to design and submit to measures that improve the successful operation of their first moral power, both in its *epistemic*

power.” He then contrasts this to his own approach to stability. My approach constitutes a promising middle ground, according to which stability *is* a task—like Rawls’s description just quoted—but a task we undertake ourselves, to shore up our own support for justice. *Political Liberalism*, p. 142. That is not to say there is never a place for engaging unbelievers, a task to which I turn in Chapter Six.

⁹⁴ Estlund, “Reply to Copp, Gaus, Richardson, and Edmundson,” *Ethics* 121, 2 (2011): 354–89, p. 358.

function—identifying what justice requires—and its *motivational* function—successfully regulating conduct. This is only part of a stability charter for a well-ordered liberal democracy, oriented around the duties each citizen has to fortify *herself*. Later, in Chapter Six, I shall explain how to think about certain duties we have to fortify others. Moreover, the stability charter of a well-ordered society is only one element of the broader *theory of stability*: the general theory of how to improve the operation of citizens' first moral power.

§4. Fortificational Duties

4.1. The self-fortification thesis

The self-fortification thesis is the contention that

Citizens are subject to moral requirements to create and submit to measures that increase the likelihood that they will accurately identify and freely affirm reasonable norms of justice.

According to this thesis, citizens of a well-ordered society are subject to moral requirements to achieve or sustain the basically just character of their society by creating, sustaining, and submitting to civic practices that stabilise their first moral power against predictable threats to its successful operation.

To pump our intuitions in a way that will, I hope, dispose us to be sympathetic to this thesis, consider a contrived example:

Tumour 1. The doctor has told you that you have a benign tumour in your brain that exerts an effect on your practical reasoning, tempting you to choose to act wrongly when you perceive that doing so will serve your narrow, immediate self-interest. The tumour was not caused by any choice or series of choices for which you are responsible. It is inoperable—it is here to stay and is therefore a permanent feature of your mental constitution. However, you can mitigate its effects through various exercises that strengthen the countervailing inhibitions against acting wrongly.

While the case is peculiar,⁹⁵ and I do not assert that it is a fully analogous case—an assertion that would commit me to controversial theses about identity and its relation to responsibility—I suspect that many people would have the intuition in this case that one ought—subject to certain conditions, to be discussed later, about the content and demandingness of the exercises—to undertake the exercises. And the reason why one ought to perform the exercises is because we have decisive moral reason to respect our fellow citizens’ moral status. By failing to take proactive action against the potential effects of the tumour, we expose our fellow persons to considerable risk of wrongdoing. Call the process of combatting the tumour in *Tumour 1* the process of *motivational fortification*.

Consider now a different case:

Tumour 2. The doctor has told you that you have a benign tumour in your brain that has a tendency to cause you to reason defectively about what justice requires in particular cases—making fallacious inferences, endorsing morally relevant empirical claims without sufficient evidence, embracing conclusions that do not follow deductively from the preceding premises, mis-identifying the weights of different reasons for action, etc. The tumour was not caused by any choice or series of choices for which you are responsible. It is inoperable—it is here to stay and is therefore a permanent feature of your mental constitution. However, you can mitigate its effects through various exercises that strengthen your countervailing tendencies toward sound moral reasoning.

I also suspect that many people would be inclined to believe that—subject to certain conditions—one ought to undertake these exercises. For lack of a better term,⁹⁶ call the

⁹⁵ For some scepticism on using outlandish cases, see Jakob Elster, “How Outlandish Can Imaginary Cases Be?” *Journal of Applied Philosophy* 28, 3 (2011): 241-58.

⁹⁶ This is a non-ideal term since it appears metaethically partisan—suggesting that justified claims about justice are facts about the world that the mind *discovers*, as opposed to upshots of practical reasoning that the mind *constructs* (or that minds *together* construct, as an intersubjective creation). I use the term in a manner compatible with all such cognitivist positions. A potentially superior term would be *cognitive stabilisation*. However, those who endorse certain anti-Humean positions in moral psychology think that motivation is *itself* best understood cognitively; on such a view, both categories I am discussing would belong to a category termed cognitive stabilisation. To avoid these confusions, I stick to the terminology I have offered here.

process of combatting the influence of the tumour in *Tumour 2* the process of *epistemic fortification*.

The self-fortification thesis's plausibility does not hang on these toy examples; its plausibility hangs on the way the idea of stability, as I understand it, makes sense of elements of our political practice to which we are reflectively committed. The tumours in these examples are, obviously, representative of human epistemic and motivational fallibility, as that fallibility is confirmed by the phenomenology of agential experience, the empirical literature on human moral fallibility, and serious reflection on contemporary politics and human history. My conviction is that we—living, breathing, embodied, socialised, fallible human agents—find ourselves in a situation sufficiently similar to that of the agents in *Tumour 1* and *Tumour 2* that, *mutatis mutandis*, similar conclusions follow. Our capacity for justice—our moral power to act in ways that respect others as free and equal—is fragile. This fragility leaves us with two general options: either we do something to mitigate that fragility, or we do not. To understate my thesis: justice is not indifferent toward which we choose.

Let us review our intuitions on the two cases I have described to ensure they are not anomalous, but instead rest in equilibrium with our considered judgements. We have reason to reflectively endorse the intuitions in the *Tumour* cases once we see their place in a larger body of considered judgements we have, firstly, about *negligence* and *endangerment*, and, secondly, about *personal responsibility for moral duties*. I take each in turn. Consider the example of extreme speeding in one's automobile. The reason that extreme speeding is wrongful is not that it violates anyone's substantive rights, narrowly conceived.⁹⁷ The reason it is wrongful is that it exposes others to the *risk* that their rights will be violated.

⁹⁷ That is, conceived independently of the conviction I am about to articulate.

It endangers them without sufficient justification.⁹⁸ Now consider a case before the New York Court of Appeals in 1956. The defendant, Emil Decina, was an epileptic. Knowing he was prone to debilitating seizures, he nevertheless chose to drive a car. He experienced a seizure while driving, and the car swerved. It killed four young girls. The court considered the question of whether Decina's conduct qualified as culpable negligence. The judge who authored the majority opinion noted that while "to have a sudden sleeping spell, an expected heart or other disabling attack, without any prior knowledge or warning thereof, is an altogether different situation,"⁹⁹ the fact that Decina knew he was prone to epileptic fits made him criminally negligent—simply by *choosing* to get into the car and drive it. Setting aside the legal questions of the particular case, this result strikes us as morally sound: people have a duty not to act in ways that pose dangers to others without sufficient justification.

Now these two cases are partially disanalogous from the state of affairs in which, I am suggesting, all of humanity finds itself. What makes speeding so bad is that the driver's decision to speed places her in a state of affairs in which she may be incapable (on the physicalist reading of "ought implies can") of avoiding the violation of someone's right (e.g., because she will be going too fast to swerve out of the way of an innocent pedestrian in time). That incapacity is even more striking in the case of the knowingly epileptic driver. I have designed the *Tumour* cases, however, such that the agent necessarily takes herself to *retain* the capacity to choose rightly when a moment of potential rights-violation arrives; what matters is that she is simply *unlikely* to exercise it without having done the preparatory work. The *Tumour* cases are thus designed to mirror a crucial feature of our agential phenomenology for which I argued earlier: it belongs to

⁹⁸ For discussion on risk, see Jules Coleman, *Risks and Wrongs* (Cambridge: Cambridge University Press, 1992).

⁹⁹ *People v. Decina*, 1 A.2d 592 (N.Y. Ct. App. 1956), Section 3.

our self-understanding to see ourselves as capable of choosing to live up to moral demands. But there is a worry that this very feature renders our considered judgements in the driving cases irrelevant to the *Tumour* cases. For, an agent might complain: “I understand why it would be wrong to *remove from my power* the choice to respect others’ rights. But so long as I retain my power—and as you say, as an agent, that is what I must presuppose myself to do—why should I have to engage in the these tiresome fortificational practices? Trust me: when the moment of vulnerability comes, I will do what is right, no-fortification-necessary.” But the exhortation to trust is not credible. This is partly because we can modify the driving examples to weaken the disanalogy (e.g., we can consider the case of *moderate* speeding as opposed to extreme speeding, which is wrongful only because it makes the driver *less likely* to avoid a rights-violation when a potential moment comes, not *incapable*). The central reason the arrogant agent’s exhortation to trust should not be taken seriously is simply what we know about human fallibility. The proposition that people’s temptations and circumstances can render them less likely to choose to act rightly is not remotely controversial, and the burden of proof falls on an agent who *denies* her susceptibility to such countervailing forces.¹⁰⁰

The second body of convictions that bears on our reflective endorsement of the intuitions in the *Tumour* cases concerns our convictions about agents’ *personal responsibility for moral duties*.¹⁰¹ When we say that Barbara is subject to a duty not to engage in murder, this is not some obtuse shorthand for saying that *someone else* has to see to it that Barbara does not engage in murder. It means that *Barbara* needs to see to it that she does not murder anybody. This helps us respond to an agent who offers the following objection:

¹⁰⁰ Anyone in a contemporary liberal democracy who *denies* this neglects the extent to which her psychological constitution has *already* been subjected continually to the numerous fortificational pressures that belong (though not nearly as centrally as they ought to) to our moral and civic life.

¹⁰¹ See the distinction between responsibility as attribution and substantive responsibility in T. M. Scanlon, *What We Owe to Each Other*, p. 248.

“But I never *asked* to be fallible; I never *chose* to be placed in a situation in which I posed a risk to the rights of others. Accordingly, I am not responsible for undertaking these putative fortificational measures.” But this response defies our normative self-understanding. Our objector’s complaint is similar to that of someone who passes a drowning child and says: “I never *chose* to be the only person in this child’s proximity.” While a person’s opportunity to have avoided certain imposed burdens can sometimes be relevant to the justification of imposing them upon her,¹⁰² it is a profound misuse—and misunderstanding—of this idea to suggest that any morally assigned burden that has not been freely consented therefore cannot be a morally assigned burden. *That* determination depends on whether there is a rival normative principle that does not expose the agent to the burden in question but is nevertheless morally plausible. But no such principle can be found without upending many of our most fundamental normative convictions. Surely the objector would not insist that *other people* are substantively responsible for increasing the likelihood that she not violate their rights (by, for example, staying out of her way)? This is so *even if* it is difficult for her to respect their rights because, for example, she takes extraordinary pleasure from harming others. On this point, I follow Rawls:

[T]he principles of right and justice are collectively rational; and it is in the interest of each that everyone else should comply with just arrangements...Those who find that being disposed to act justly is not a good for them cannot deny these contentions. It is, of course, true that in their case just arrangements do not fully answer to their nature, and therefore, other things equal, they will be less happy than they would be if they could affirm their sense of justice. But here one can only say: their nature is their misfortune.¹⁰³

I thus conclude that we have no reason to doubt our intuitions in the *Tumour* cases, and thus to accept the intuitive pressure they adduce in favour of my self-fortification thesis.

¹⁰² *Ibid.*, p. 257ff.

¹⁰³ Rawls, *A Theory of Justice*, p. 504. Rawls’s point applies with equal force to *Tumour 2* as it does to *1*; a racist who has recently renounced her past beliefs must not say that she was not actually *subject* to the moral requirement not to be a racist.

Note two distinct implications of the self-fortification thesis. One possible implication is that it offers grounds for believing that citizens *here and now*, in presently far-from-just societies, ought to fortify our moral powers against countervailing pressures. The second implication is *even citizens of basically just societies* would have reason to endorse stabilising mechanisms. I believe that both implications follow. However, I restrict myself here to exploring the second implication. I do so because an adequate understanding of the second implication is a necessary prerequisite for understanding the first. We need an *ideal* of stability—an ideal of how morally motivated but fallible citizens could and should work together through well-functioning democratic institutions to stabilise their commitment to justice against countervailing forces—before we can turn to the theoretically messier task of how such stabilisation could or should proceed in the context of *systematically malfunctioning* institutions of the kind that characterise an unjust society. That those malfunctioning institutions themselves *exacerbate* tendencies toward wrongdoing—serving to destabilise people’s moral powers at both the levels of motivation and reasoning—raises serious questions about, for example, their fitness for serving as instruments of stabilisation (both in terms of their pragmatic effectiveness and their “moral standing”¹⁰⁴). The aim of this dissertation is narrower, focusing primarily on the question of what stabilising measures might be necessary to *sustain* a well-ordered society. With that picture in hand, future work can subsequently attend productively to the way that the stabilising measures I theorise could also be harnessed to *transition* from an unjust society to a well-ordered liberal one. We should not be surprised if the civic practices that help us *sustain* a just order bear *some* relation to the civic practices we have

¹⁰⁴ See, for example, Jeffrey Howard, “Punishment, Socially Deprived Offenders, and Democratic Community,” *Criminal Law and Philosophy* 7, 1 (2013): 121-136.

reason here and now to create and sustain as part of a moral mission to *achieve* a just a order.

A few more comments are in order. Firstly, we can distinguish between *fortificational measures* and *fortificational duties*. *Fortificational measures* empower people to exercise their first moral power more successfully. Such practices can be institutional mechanisms, civic practices, policies, social norms, or personal norms. *Fortificational duties* are the moral requirements to which we are subject to *create* the measures, *sustain* them, and *submit* to them. To justify a fortificational measure just *is* to justify the claim that citizens are subject to duties to create, sustain, and, under certain conditions, submit to such a measure. Note that some fortificational measures are created through formal legal enactments, others can be informally pursued throughout civil society.

Second, we can distinguish between *general fortificational duties* and *special fortificational duties*. *General fortificational duties* are the duties citizens have, simply in virtue of their fallibility, to design or otherwise sustain particular fortificational measures. *Special fortificational duties* are the duties citizens have to *submit* to the particular measures. These duties are *special* in the sense that their particular content may vary from citizen to citizen, depending on the duty and the case. Some fortificational measures aim simply to shore up people's sense of justice *simpliciter*, whereas others aim at fortification with respect to some particular subset of fundamental duties of justice. So, while all citizens are subject to a general fortificational duty to create and maintain a scheme of publicly funded counselling services—since no one can guarantee in advance that they will not need to avail themselves of the service—people will have varying special duties to avail themselves of it. A citizen who experiences temptations to lash out violently at others has a special fortificational duty to avail herself of the counselling, whereas someone who experiences no such temptations does not. The content of one's special fortificational

duties consists in what they could reasonably be expected to know about their own degree of susceptibility to temptation or error with respect the relevant issue. Given the importance of retaining the moral quality of social relationships, special fortificational duties are not coercively enforced. In certain non-ideal cases, however, coercion can be appropriate—albeit regrettable—which I shall address in Chapter Five.

Finally, while duties of self-fortification are best understood as duties that each citizen has in virtue of her own fallibility, in many cases citizens can most effectively pursue fortificational tasks by working together to sustain them collectively. This is especially important in the case of epistemic fortification, in which citizens are attempting to improve the effectiveness of their moral reasoning, but it applies to motivational fortification, too. Moreover, while I remain agnostic for the purposes of my analysis here on the proper method of *funding* fortificational measures that require funding, the fact that all citizens are fallible and that all therefore properly avail themselves of different features of the fortificational infrastructure over the course of their complete lives makes it not unreasonable to adopt a simplifying assumption that such measures are to be funded jointly.

4.2. Epistemic self-fortification: the improvement of moral reasoning

Our sense of justice enjoins us to do two things: firstly, to *figure out* what justice requires us to do; and secondly, to *do* it. Let us take each in turn.

Epistemic fortification is oriented to the task of increasing the reasonableness of people's moral convictions. That such epistemic stabilisation will be necessary even in a well-ordered society is evident from reflection on two important facts. Firstly, even if most citizens of a liberal society have been socialised to endorse liberal norms of justice, human moral reasoning remains susceptible to the biases and prejudices known to social

psychology.¹⁰⁵ Secondly, fresh concerns confront human affairs even in the well-ordered society. Our developing understanding of natural and social science—and the attendant development of fresh technologies—leads us to encounter new, high-stakes questions whose answers do not simply drop out of prevailing theories of justice.

The theory of epistemic fortification concerns practices, mechanisms, or norms justified by their function in increasing the intelligence of citizens' moral reasoning. Such practices may be freestanding, or they may be institutionally integrated into independently justified practices, harnessing elements of them in ways that invite sensible reflection. *Democratic discourse* serves as among the most important stabilising forces: the existence of a critical culture in which citizens talk to one another about the justifiability of their convictions. It is not a single practice, but comprises manifold particular norms, both informal and formal. Some informal norms include: the norm of criticising a friend or colleague when she says something with which one disagrees; the norm of discussing questions of justice around one's dinner table; the norm of reading the news, and writing letters to the editor when one sees something one disputes; the norm of going out of one's way to read a columnist who affirms a conception of justice one disputes, given the empirical evidence of confirmation bias; the practice of hosting a community discussion group on matters of common concern and continually inviting fellow citizens with whom one disagrees on many issues, so as to resist the effects of in-group polarisation;¹⁰⁶ and so on. Some formal practices of democratic discourse include:¹⁰⁷ the practice of conducting and broadcasting the results of "deliberative polls," in which diverse

¹⁰⁵ I discuss these at the beginning of Chapter Three.

¹⁰⁶ Cass Sunstein, *Going to Extremes: How Like Minds Unite and Divide* (New York: Oxford University Press, 2009).

¹⁰⁷ These examples in the next three footnotes are compiled in Robert E. Goodin and John S. Dryzek, "Deliberative Impacts: Macro-Political Uptake of Mini-Publics," *Politics & Society* 34, 2 (2006): 219-244.

constituencies of citizens discuss matters of common concern with experts, with before-and-after comparisons of participants' beliefs made available;¹⁰⁸ the official creation of "deliberation days" to occur prior to national elections;¹⁰⁹ public consultations in which governments organise large groups of citizens to consider evidence and issue recommendations; and "citizens' juries" of the kinds held in the United Kingdom, the United States, Canada, Denmark, and Germany, in which governments convene small groups of citizens to discuss or generate particular proposals.¹¹⁰ Given citizens' duties to implement reasonable policies on various matters of common concern, and given their knowledge of their epistemic fallibility, they have duties to set up and pursue special processes that make them more likely to arrive at reasonable results on the particular policy questions.

Part of my aim in the following chapters is to help cast some features of our political practice as epistemically fortificational measures. Firstly, I will examine the practice of mainstream democratic discourse. Secondly, epistemic stabilisation will figure in my discussion of constitutional design, particularly in relation to judicial review's didactic functions. Note two further issues that relate to epistemic fortification, but are

¹⁰⁸ James S. Fishkin, *Democracy and Deliberation: New Directions for Democratic Reform* (New Haven: Yale University Press, 1991); and *The Voice of the People: Public Opinion and Democracy*, revised edition (New Haven: Yale University Press, 1997).

¹⁰⁹ Bruce Ackerman and James S. Fishkin, *Deliberation Day* (New Haven: Yale University Press, 2004).

¹¹⁰ For details, see Jefferson Center, "Jefferson Center: Originator of the Citizens Jury Process," <http://www.jefferson-center.org> (accessed 11 March 2013); Anna Coote and Jo Lenaghan, *Citizens' Juries: Theory into Practice* (London: Institute for Public Policy Research, 1997); and Jo Lenaghan, Bill New, and Elizabeth Mitchell, "Setting Priorities: Is There a Role for Citizens' Juries?" *British Medical Journal* 312, 7046 (1996): 1591-4; Simon Joss and John Durand (eds.), *Public Participation in Science: The Role of Consensus Conferences in Europe* (London: Science Museum, 1995); Richard E. Sclove, "Town Meetings on Technology: Consensus Conferences," <http://www.loka.org/pubs/techrev.htm> (accessed 11 March 2013); "Planungszelle & Bürgergutachten," <http://www.planungszelle.de> (accessed 11 March 2013). For some scepticism on the roadblocks to the feasible implementation of citizens' juries, see Damien French and Michael Laver, "Participation Bias, Durable Opinion Shifts and Sabotage through Withdrawal in Citizens' Juries," *Political Studies* 57, 2 (2009): 422-50.

not properly justified under the umbrella of the self-fortification thesis. The first concerns duties to engage with unreasonable citizens, which I defend through the natural duty to support just institutions in Chapter Six. The second concerns schoolhouse civic education; this issue, at least for ideal societies, has been treated extensively in the scholarly literature and I largely agree with what many Rawlsian liberals have already said on the matter.¹¹¹ It is relatively uncontroversial that we have duties to nourish the first moral power of children. Far less discussed is the proposition that adults have duties to nourish *our own* first moral power.

4.3. Motivational self-fortification: improving moral decision-making

Motivational self-fortification aims to close the gap between human motivational sensibilities and what is morally required. Some examples of motivational fortification work by strengthening agents' moral dispositions, others by reducing the influence of countervailing dispositions. Some strategies are *agent-centric*—proceeding through forms of education, training, or therapy. Others are *environmental*—proceeding through the

¹¹¹ See especially Stephen Macedo, *Diversity and Distrust: Civic Education in a Multicultural Democracy* (Cambridge, MA: Harvard University Press, 2000); Martha Nussbaum, *Cultivating Humanity* (Cambridge, MA: Harvard University Press, 1997), and *Not For Profit: Why Democracy Needs the Humanities* (Princeton: Princeton University Press, 2010); and Amy Gutmann, *Democratic Education* (Princeton: Princeton University Press, 1987). See also Eamonn Callan, *Creating Citizens* (New York: Oxford University Press, 1997); William Galston, *Liberal Purposes* (Cambridge: Cambridge University Press, 1991), pp. 241-256; and Meira Levinson, *The Demands of Liberal Education* (New York: Oxford University Press, 1999). These latter thinkers—Callan, Levinson, and Galston—insist that civic education theory demonstrates the untenability of the distinction between political and comprehensive liberalism, since teaching students about the burdens of judgement inevitably inculcates them to accept an autonomy championing comprehensive doctrine. I am sceptical of that; while Rawls does recognise the worry (*Political Liberalism*, p. 200), it remains true that (1) recognition of the difficulties involves in coming to agreement on many important questions, and (2) affirmation of the idea that an object's or project's value is traceable only to the fact that it was willed by an autonomous first-mover, are clearly not the same thing. One can accept the burdens of judgement and wholeheartedly reject the ideal of autonomy—indeed, even reject the ideal of free and equal citizenship. For important discussion on the different ways of coming to understand the burdens of judgement, see John Tomasi, "Civic Education and Ethical Subservience: From *Mozart* to *Santa Fe* and Beyond," in Stephen Macedo and Yael Tamir (eds.), *NOMOS XLIII: Moral and Political Education* (New York: New York University Press, 2002).

design of arrangements that call forth the expression of certain dispositions in people and discourage the activation of others.

Epistemic fortification is easier to conceptualise than motivational fortification, especially for normative political theorists. Political theorists are, in some sense, in the business of epistemic fortification—we aspire to reduce the likelihood that we will make mistaken inferences or deductions, misidentify the weights of different reasons for action, and so on—and indeed we spend most of our time criticising others when they make such mistakes. The whole practice of education *presupposes* the possibility that, with hard work, we can discipline our thinking. But motivational fortification is problematic also because a single, master view on what it involves would require taking sides in debates on moral psychology that have raged since Plato: about the nature of the human mind; about the relation between reason and emotion in moral judgements; and about the role of desire in moral motivation. The controversy among moral psychologists about the structure and nature of moral motivation is vast and deep. Humeans point to the inescapable role of natural passions in explaining moral motivation; Kantians point to the inescapable role of reason in *subordinating* our natural passions in explaining moral motivation; Aristotelians point to the development of character, cultivated through habituation and the consequential acquisition of virtue; and still others dispute all of them, insisting that any human being can be brought to act with unimaginable heroism, or unfathomable evil, simply by placing her in the right context.¹¹²

The problem is that moral psychologists *reasonably disagree* about what, exactly, moral motivation *is*. They disagree about its nature, and thus its existence conditions. Accordingly, they disagree about what has “gone wrong” in instances of motivational failure. I shall have occasion to consider rival views about what motivational failure

¹¹² I review these categories in greater detail in Chapter Five.

involves in greater detail in Chapter Five. For now, a vital question arises: how do we devise a normative political theory oriented to the aim of motivational fortification without a determinate understanding of what such a process precisely involves? How is *any* justified theorising about this area of the theory of stability possible in the face of such disagreement?

I wish I could tell you that I had performed some experiment, or secured access to some heretofore-unanalysed fMRI study, or developed a new technology for predicting the movement of molecules in the brain, that would enable me to arbitrate between these different factions in moral psychology. However, the fact that I have not is no embarrassment to my view. Moral philosophers, psychologists, economists, and neuroscientists continue to disagree fiercely about the nature of the human mind and the workings of rationality—both within their own camps, and across them. The key is to distinguish the kind of reasonable disagreement that is at issue. Now many debates in moral psychology concern deep questions about the metaphysics of personhood, and the relation between moral motivation and underlying cosmic forces, that belong to the philosophy of mind and theology. These debates are reasonable, in the sense that the burdens of judgement ensure continuing contestation on such questions. But our interest in moral psychology is not immediately *theoretical*; we are not interested in settling debates on centuries-old questions. Our interest is *practical*. Citizens sense the fragility of their motivational capacities, and they want to know what to do about it—they want to know *enough* about what moral failure *is* that they can take proactive measures to reduce its likelihood. It will not do to say in response to such citizens: “Sorry—public policy must never be informed by any empirical claims about human motivation; they fall outside the ballpark of public reason.” If *that* were true, no claims about, for example, the likelihood of agents responding to prudential incentives could ever be a part of public policy, for it

is a claim about human motivation; significant swaths of public policy would automatically become unjustified. But that cannot be right.

The real reasonable disagreement about moral psychology that matters for political liberals is not the deep theoretical disagreements about whether our temptations were implanted by Satan or any other matters that citizens are free to work out for themselves. Rather, what matters are the practical reasonable disagreements concerning different actionable proposals for motivational fortification—all justified on the basis of historical reflection and evidence from the social sciences, *both* of which are public reason-consistent,¹¹³ not on the basis of controversial metaphysical or theological premises (though each proposal would be compatible with various deeper stories). Indeed, the core insights of each view I mentioned—Humean, Kantian, Aristotelian, or situationist—each enjoy considerable empirical and phenomenological corroboration, and are compatible with many underlying views (and, in places, one another) when untethered from their metaphysical roots. Given the shared conviction that *some* fortificational policies ought to be adopted, it is enough simply to say: so long as a proposed policy is reasonable (an issue I confront in the next section), it does not disrespect others for the political community to authorise it as a morally required component of the official stability charter. The same practice that applies to our

¹¹³ I follow Rawls in making this point (*A Theory of Justice*, p. 175). There is significant unexplored work to be done in political philosophy by Rawlsian liberals about what sorts of social-scientific claims—and natural-scientific claims—belong to public reason and which do not. My conviction is that people who believe that facts secured from biology, physics, and chemistry cannot be used in premises of arguments for public policy are unreasonable. Because the natural sciences are more likely to implicate deep questions about the “nature of the universe”—a favourite topic of comprehensive doctrines—their admission into public reason *a fortiori* would appear to vindicate social-scientific claims, which concern, *inter alia*, predictive claims (not iron-clad laws of nature) about how people are likely to behave under various conditions. Gaussian defenders of public reason, who index their understanding of public reason’s domain to idealised variations of the belief sets of the relevant constituency, this is a somewhat more tractable problem. For people like Rawls himself—alongside Charles Larmore, Joshua Cohen, and Jonathan Quong—this is a far trickier matter that I cannot pursue here.

reasonable disagreements about scientific and economic policy—settling them democratically—applies to our reasonable disagreements about moral psychology.

There are several important areas in the theory of motivational fortification. Firstly, it concerns the question of how to reduce the likelihood that people will choose to engage in crime—a topic I take up in Chapter Five. Secondly, it concerns how to reduce the likelihood that voters and officials will choose knowingly to enact unjust legislation—a topic I take up, as a case-study in the theory of institutional design, in Chapter Four. Thirdly, it concerns a wide host of regulatory issues about how to design the contexts in which individuals make choices such that they are more disposed to select the rightful course of action, thereby “nudging” people to make good decisions.¹¹⁴

§5. Public Justification and Limits

This moral requirement remains considerably abstract. There are many proposals for what sound epistemic and motivational fortificational practices might involve. How are we to identify which practices are those that we, as democratic citizens, have decisive moral reason to pursue? The way to identify whether a proposed fortificational requirement is justified is the same way to identify whether *any* proposed moral requirement is justified: ask whether free and equal citizens committed to one another’s moral status would agree to its place in the enduring infrastructure of a well-ordered liberal society. Rawls himself argues that parties to the original position would agree to

¹¹⁴ Cass Sunstein and Richard Thaler, *Nudge* (London: Penguin Books, 2008). This is by far one of the most important and promising areas in the theory of stability. It is my regret that I do not take it up in any detail in the chapters that follow. It is a crucial ground of possible interdisciplinary research in the theory of stability between normative political theorists, political psychologists, moral psychologists, public policy designers, lawyers, and behavioural economists. Because engaging the issue of what Sunstein and Thaler call “choice architecture” would take me too far afield here, due to the empirical evidence it would require me to engage and evaluate, I save that project for another day.

penal mechanisms to cool citizens' destabilising suspicions that others would not be doing their part. I will have some critical comments on that particular proposal in Chapter Five, but the general idea is promising: we can envision parties to the original position, aware of the general facts of history, psychology, and social theory (as Rawls says they are)—and thus human fallibility—who take up the task of designing fortificational arrangements. We can, in other words, imagine a revised Stage 2 of the original position—one dedicated not to *testing* the stability of the principles generated in Stage 1, but in *planning* for it. And while we will inevitably disagree reasonably about what would ensure, so long as proposal is *reasonable*, its democratic authorisation triggers a duty to pursue it. But what does it mean for a proposal to be reasonable?

We can think already of three necessary criteria. Firstly, there must actually be a problem that the proposal is seeking to solve. We must have evidence that human powers of practical reason are liable to operate defectively in the way the proposal presupposes. Call this the *defectiveness criterion*. Secondly, we must have some evidence that the proposed strategy must stand some reasonable chance of accomplishing its intended aim if pursued. Call this the *effectiveness criterion*. Thirdly, given that fortificational practices can be demanding—they can require our time and energy to undertake—a proposal's demandingness should be no greater nor less than what could reasonably be regarded as standing a sufficient chance of achieving its purpose. This criterion is deliberately vague, and I shall have occasion to return to it later; call it the *proportionate burdensomeness criterion*. Thus if there is no evidence that the proposal addresses a real problem, or addresses it with any prospect of success, then it is an unreasonable proposal. What else disqualifies a stability proposal from the set of the reasonable?

Recall the *DRONES* proposal we considered earlier. Consider also: the injection of a serum into all citizens' brains that induces them to commit suicide upon forming the

intention of violating others' rights, thereby eliminating the morally weak from the gene pool; the public execution and limb-amputation of those who perpetrate serious crimes, whose corpses and body parts are left to hang in the streets to terrorise the population into complying with the law; the establishment of an official public educational scheme that teaches all children that they will suffer unspeakable pain for eternity unless they comply with the law while in this world; the insertion of a chip into everyone's mind that would activate every time the agent engaged in moral decision-making, temporarily relieving the agent of command and granting it to a state-run super-computer, which would return agential control to the agent after the period of moral reasoning and action was completed.¹¹⁵ Surely *these* also qualify as unreasonable fortificational measures, right?

It is vital to see that what disqualifies these mechanisms is *not* that they are fortificational practices that *go too far*; it is that they are not properly classified as fortificational practices *at all*. Fortificational practices are oriented to improving the likelihood that people will choose to comply freely with correctly identified fundamental norms of justice. Any norm that does not have the fortification of the sense of justice as its aim simply is not a fortificational norm. This rules out—as candidates for fortificational norms, anyway—the dramatic proposals I have just listed, which serve not to *engage* and *empower* our moral agency, but *bypass* it: terrorising or manipulating us into compliance. Fortificational norms serve to increase the likelihood of stability *for the right reasons*. But choosing not to affront others' rights because we fear the consequences of doing so is the *paradigmatic* instance of “acting justly” for the wrong reasons. It defies our ideal of just social relationships among fellow citizens—the defining ideal of the well-ordered society. This applies not merely to the coercive proposals, but also to the

¹¹⁵ Supposing, implausibly, we could isolate periods of moral decision-making from non-moral decision-making.

deceptive ones. Plato's proposal that philosopher-kings lie to the masses about their nature cannot be conceived as a fortification of those masses' moral agency, for it co-opts and undermines that agency.¹¹⁶

Thus our limited objection here to the terrorising proposals above is not that they are *bad* fortificational proposals; it is that they are not fortificational proposals. They do not belong to our ideal of a well-ordered society of fallible human beings who nevertheless work hard to give justice freely to one another.¹¹⁷ This helps us rule out other proposals that some might mistakenly regard as fortificational norms. Consider proposals that restrict prospective victims' freedoms for the sake of reducing the quantity of rights-violations. For example, we might imagine a proposed requirement enjoining women "not to walk alone in certain areas at night, or not to wear certain kinds of 'provocative' clothes in certain places," so as to assist others in choosing not to mug or rape. Or imagine a proposed requirement instructing "black men...not to visit certain pubs which are known to be frequented by racists."¹¹⁸ Both proposals are morally pernicious. Fortification is about *shoring up* our moral powers to make possible a social world in which we freely live together as free and equal, despite potentially pernicious tendencies within our nature. But such proposals *accommodate* those pernicious tendencies,

¹¹⁶ Rawls mentions a similar possibility: "[S]uch devices as Plato's Noble Lie in the *Republic*, bk. III, 414–415, are ruled out, as well as the advocacy of religion (when not believed) to buttress a social system that could not otherwise survive, as by the Grand Inquisitor in Dostoevsky's *The Brothers Karamazov*." Rawls, *A Theory of Justice*, p. 398n.

¹¹⁷ I thus focus on one small area of what we might call "wrongdoing prevention." That is a massive and important topic, but I am focused in this dissertation only on a small set of examples from one important slice of that topic: preventing wrongdoing through fortifying citizens' first moral power. For recent work on the wider debate, see Andrew Ashworth, Lucia Zedner, and Patrick Tomlin (eds.), *Prevention and the Criminal Law* (New York: Oxford University Press, 2013). See also Brandon C. Welsh and David O. Farrington (eds.), *The Oxford Handbook of Crime Prevention* (New York: Oxford University Press, 2012).

¹¹⁸ R. A. Duff and S. E. Marshall, "Benefits, Burdens and Responsibilities: Some Ethical Dimensions of Situational Crime Prevention," in *Ethical and Social Perspectives on Situational Crime Prevention*, edited by Andrew von Hirsch, David Garland and Alison Wakefield (Portland, OR: Hart Publishing, 2000), 17–35, p. 27.

imperilling the achievement of a just social world in the process. Fortificational measures aim instead to *reckon* with our fallibility, not make room for it to reign free. This does not mean that duties of stability do not involve costs, demanding our time and energy. But those costs are justified because of their role in sustaining our relations as free and equal; they are not costs *to* justice. The broader lesson is this: because our aspiration is to *sustain* our just social relationships, we would not endorse mechanisms whose character would destroy the moral quality of those very relationships. Fortificational measures by their nature cannot do this, lest they no longer *be* fortificational.

Conclusion

Rawls famously divides the project of normative political theory into two parts: *ideal theory* and *non-ideal theory*. The task of ideal theory—or, more precisely, the task of constructing an ideal theory *of justice*¹¹⁹—is the task of determining a vision of an ideal social order in a manner that adheres to Rousseau’s directive: “taking men as they are and laws as they might be.”¹²⁰ A vision so specified, Rawls writes, constitutes our “realistic utopia,” the vision of society toward which we ought to strive. It is not an exercise in “the ethics of creation”¹²¹—inventing the laws of physics and biology—but rather it asks what we can and should aim for, given the kinds of beings we are. *Non-ideal theory*, on the other hand, concerns how we should respond to injustice, to severe natural limitations (e.g., children and the mentally ill), and to unfavourable cultural conditions, as well as how we should transition to a just society from an unjust one. To what does a concern with stability, we are now in position to ask, belong?

¹¹⁹ Patrick Tomlin has persuasively argued that “ideal theory” and “non-ideal theory” are unhelpful terms unless indexed to particular ideals. See “Should We Be Utopophobes About Democracy in Particular?” *Political Studies Review* 10, 1 (2012): 36-47.

¹²⁰ Rousseau, “On the Social Contract,” p. 141.

¹²¹ Rawls, *A Theory of Justice*, p. 137. See also Miller, “Political Philosophy for Earthlings.”

Rawls avows that the concern of how a well-ordered society would remain stable is one belonging to ideal theory—our theory of what just social relationships properly involve, given human beings as they are. I believe he is right about that. But this chapter has endeavoured to show that the *way* this concern belongs to ideal theory is very different than what Rawls invites us to believe. Stability concerns do not operate as a *constraint* on the content of our normative convictions—forcing the philosopher quietly to dilute her aspirations before she delivers their content to the world, like a director who must cut some of the best-written scenes of the play minutes before the curtain opens because the actors do not seem quite up to the task. Rather, the concern with stability is a practical problem that citizens of the well-ordered society are required by justice to solve, by designing and submitting to measures that increase the intelligence of their moral reasoning and motivational resilience of their sense of justice. The question of how it is that we can undertake just social relationships that endure over time is of paramount importance, and the aspiration to find and implement some reasonable answer belongs to the shared ideals that characterise reasonable citizens’ common ground in the ideal society—and thus should belong to *our* ideal vision. I have thus specified a *final* element of the commitment to justice that ought to belong to an overlapping consensus: an aspiration to *fortify* the rest of the ideals, despite reasonable disagreements about how such fortification is to be done. That justice enjoins us to pursue such fortification need not reflect any judgement that our human nature is defective or lamentable (though it is not incompatible with such a judgement). Instead it reflects the plain truth that justice is hard work.

The question of how we will make the freely willed achievement of justice an enduring reality, given the fallibility of our nature, is the basis of a compelling, empirically engaged research agenda in political philosophy. But our understanding of stability is not

merely important because it helps frame political proposals for new practices and institutions. As I shall now attempt to show, it helps us understand the best rationale for many *prevailing* practices and institutions, revealing stability's under-acknowledged role at the heart of liberal-democratic political and moral life.

CHAPTER THREE

*Deliberative Democracy and the Fortification of Moral Reasoning*¹

§1. Pitfalls of Moral Reasoning

Everyone thinks they know injustice when they see it. We seethe with anger when we hear the outcome of a court ruling that all but guarantees corruption in political campaigns. We slam our fists on the kitchen table in rage upon reading some newspaper article about an unjustified war fought in our name. We curse at the television upon seeing corrupt officials who promise equality of opportunity and social mobility in their speeches while systematically rigging the terms of social interaction against those unlucky enough to be born poor. And we experience a saddening alienation from our fellow citizens when we learn that they have voted for some pernicious reform that strips a vulnerable minority of a civil right.

We also feel we know justice when we see it. We cheer with happiness when some flagrantly unconstitutional provision is struck down. We cry with joy when we see previously humiliated and disenfranchised citizens affirmed and empowered by a legislative decision. We exhale a sigh of relief when we learn that some torture centre is going to be closed. We slap our open palms against our fellow citizens' upon learning that our preferred candidate, who favours extensive educational reforms, has won an election.

We do all these things, and we do them with confidence. But what makes us so sure we are right? What is the ground of the confidence in our convictions that these reactions betray? Two insights shake our epistemic faith. The first concerns liberalism's apparent elusiveness. Human history renders laughable Thomas Jefferson's conviction

¹ This chapter will appear in late 2013 as Jeffrey Howard, "Democracy as the Search for Justice: An Epistemic Justification," *Political Studies* (forthcoming).

that the moral equality of persons is “self-evident.”² However confident we may be that liberal ideals simply drop out of pure practical reason, as Kant avowed, or the pragmatic presuppositions of communicative rationality, as Habermas insists,³ they certainly have not *struck* most of humanity as obvious. What is most alarming of all is the systematic *misunderstanding* of liberal principles exhibited by numerous people—including master thinkers. America’s original sin, slavery, is a lesson that our sense of justice has a blind spot. Even philosophers forward-thinking enough to advocate for the equality of the sexes—such as J. S. Mill—could not be relied upon to say sensible things about the people he called “barbarians.”⁴ At all turns, the concepts and categories of liberal political philosophy have been marshalled with brutal efficiency against the moral aspirations of liberalism itself. Sometimes it simply has been a matter of circumscribing the circle of public justification—cutting out the poor, or women, or Jews, or Catholics, or gays and lesbians, as though they lacked the relevant “range property.”⁵ At other times it has been more complex, misunderstanding the implications of a particular element of liberal political theory. The public/private distinction is one obvious example. The notion of a natural right to private property did the political work Locke needed it to do.⁶ But in the nineteenth and early twentieth centuries, its role in Anglo-American moral consciousness was to immunise economic life from moral scrutiny. It took a long time to loosen the grip of that understanding; G.A. Cohen even thought that Rawlsians remained beholden

² Thomas Jefferson, “The Declaration of Independence,” in *The Declaration of Independence and the Constitution of the United States* (New York: Bantam, 1998), p. 53

³ For example, Kant, “Groundwork of the Metaphysics of Morals,” although the development of Kant’s moral theory along liberal lines in the Rawlsian tradition neglects the fact that he was insistent to separate moral from political philosophy; see the instructive discussion in Arthur Ripstein, *Force and Freedom: Kant’s Legal and Political Philosophy* (Cambridge: Harvard University Press, 2009), pp. 1ff. See also Habermas, *Moral Consciousness and Communicative Action*, pp. 43-115.

⁴ J. S. Mill, “A Few Words on Non-Intervention,” in *Collected Works of John Stuart Mill: Essays on Equality, Law, and Education*, vol. XXI, edited John M. Robson (Toronto: University of Toronto Press, 1984), pp. 118ff.

⁵ Rawls, *A Theory of Justice*, p. 444ff.

⁶ Locke, *Second Treatise of Government*, pp. 18ff.

to a more subtle, and therefore perhaps more insidious, version of it.⁷ The idea that family life is immune from public inspection, except when it turns diabolical enough to be tabloid-worthy, remains a serious idea in Western popular culture. “Until quite recently,” writes Nancy Fraser, “feminists were in the minority in thinking that domestic violence against women was a matter of common concern and thus a legitimate topic of public discussion.”⁸ George Wallace defended the people’s right to “racial and cultural freedom,” which morally underwrote racial segregation.⁹ This strategy has its updated forms, as when the disenfranchisement of gay and lesbian citizens is justified on the basis of “religious freedom.”¹⁰ Likewise, the liberal right to free speech has been interpreted by judicial elites to sanction the further infection of electoral politics with “the curse of money;”¹¹ the right of citizens to defend themselves from intruders have been co-opted to block any prospect of pacifying the gun-saturated warzones that are America’s urban ghettos;¹² and the right to private property continues to be marshalled to justify levels of inequality in education and opportunity that could not be justified on the basis of any

⁷ Cohen, *Rescuing Justice and Equality*, pp. 27-86.

⁸ Nancy Fraser, “Rethinking the Public Sphere,” in *Habermas and the Public Sphere*, edited by Craig Calhoun (Cambridge: MIT Press, 1992), p. 129.

⁹ George Wallace, *1963 Inaugural Address of Governor George C. Wallace*, Alabama Department of Archives and History. Available at <http://digital.archives.alabama.gov/cdm/singleitem/collection/voices/d/2952/rec/5>; accessed 15 February 2013.

¹⁰ The familiar version of this claim is that a political community’s right to religious freedom permits it to assert its views. The less trivial version of the claim, focusing on implications for anti-discrimination law, is explored in Fredric J. Bold, Jr., “Vows to Collide: The Burgeoning Conflict between Religious Institutions and Same-Sex Marriage Antidiscrimination Laws,” *University of Pennsylvania Law Review* 158, 1 (2009): pp. 179-230.

¹¹ Rawls, “The Idea of Public Reason Revisited,” in *Political Liberalism*, p. 449. Here Rawls is following Ronald Dworkin, “The Curse of American Politics,” *New York Review of Books* (17 October 1996), but see also Dworkin, “The Decision that Threatens Democracy,” *New York Review of Books* (13 May 2010), discussing the U.S. Supreme Court case, *Citizens United v. Federal Election Commission*, 558 U.S. 310 (2010).

¹² See the case *District of Columbia v. Heller*, 554 U.S. 570 (2008), and discussion in the run-up to the decision in Cass Sunstein, “The Most Mysterious Rights,” *The New Republic* (18 November 2007).

member of the family of reasonable liberal conceptions of justice.¹³ In some cases, these formulations are the product of malice and manipulation. But in others I see no reason to doubt the sincerity of those who believe that fundamental constitutional values demand policies that political liberalism deems to be unreasonable. And even in cases in which the rhetoric has been cunningly crafted by salaried wordsmiths and pollsters, for every citizen acting on the basis of self-interest alone we can find plenty who simply take themselves to be fighting for their convictions. If they can be so confident and yet so wrong, about what might *I* be confident but gravely mistaken?

The second insight that shakes our epistemic faith comes not from moral reflection on historical and contemporary politics, but from contemporary research in social psychology and behavioural economics. Studies demonstrate serious pathologies in belief formation about what we all-things-considered have reason to do. First, *human beings jump to confident conclusions on the basis of first impressions*. We often make lasting judgements about somebody, or some issue, based on whatever first comes to mind when we initially consider it.¹⁴ Second, *human beings' positions on a question are affected considerably by whether they are in optimistic, and thus cooperative, or pessimistic, and thus reclusive, moods*. People who hear news of someone altruistically helping someone else are more likely to play the optimal Nash equilibrium in a prisoner's dilemma.¹⁵ Third, *once human beings have formed firm beliefs on a topic, they are unlikely to take counter-evidence on that topic seriously*. When people declare publicly their commitment to a particular idea or issue,

¹³ Samuel Freeman, "Illiberal Libertarians: Why Libertarianism is Not a Liberal View," *Philosophy & Public Affairs* 30, 2 (2001): 105-51.

¹⁴ E.T. Higgins, W.S. Rholes, and C.R. Jones, "Category Accessibility and Impression Formation," *Journal of Experimental Social Psychology* 13, 2 (1977): 141-154; qtd. in Stuart Sutherland, *Irrationality* (London: Pinter and Martin, 2007), p. 11; Solomon Asch, "Forming Impressions of Personality," *Journal of Abnormal and Social Psychology* 41, 3 (1946): 258-290.

¹⁵ Harvey A. Horstein, Elizabeth LaKind, Gladys Frankel, and Stella Manne, "Effects of Knowledge About Remote Social Events on Prosocial Behaviour, Social Conception, and Mood," *Journal of Personality and Social Psychology* 32, 6 (1975): 1038-1046.

they become overwhelmingly less likely to alter their opinion on it.¹⁶ Fourth, *when people who share similar biases or beliefs spend time together, especially as part of some organised group, their biases intensify and their beliefs become more extreme.*¹⁷ Numerous experiments show that when people who already share a similar position on an issue come together to discuss it, they radicalise their position. Put a bunch of French people who distrust the U.S. together to chat with one another, and they will end the chat *really* distrusting the U.S.¹⁸ This list could continue on and on, cataloguing human beings' alarming tendencies to assume that authoritative directives are acceptable,¹⁹ to conform to group opinion,²⁰ and adopt beliefs that reinforce the status quo when they fear change.²¹ Taken together, discoveries like these render unsurprising the degree of error we see throughout historical and contemporary political life.

The question that motivates this chapter is: are the reasonable citizens of a well-ordered liberal democracy immune from the tendencies for defective reasoning so evident in historical and contemporary political life? There are powerful reasons to resist thinking that our ideal society is one in which normative reflection has stopped, replaced by a "Justice Accomplished!" banner and the entrenchment of complacency.²² The ideal

¹⁶ C. Kiesler, P. Mathog, P. Pool, and R. Howenstein, "Commitment and the Boomerang Effect: A Field Study," in *The Psychology of Commitment*, edited by Charles Kiesler (New York: Academic Press, 1971).

¹⁷ Cass Sunstein, *Going to Extremes: How Like Minds Unite and Divide* (New York: Oxford University Press, 2009).

¹⁸ Serge Moscovici, "Toward a Theory of Conversion Behavior," in *Advances in Experimental Social Psychology*, Volume 13, edited by Leonard Berkowitz (New York: Academic Press, 1980).

¹⁹ Stanley Milgram, *Obedience to Authority: An Experimental View* (New York: Harper and Row, 1974).

²⁰ Solomon Asch, "Studies of Independence and Conformity: A Minority of One Against a Unanimous Majority," *Psychological Monographs* 70, 9 (1956): 1-70.

²¹ Joseph LeDoux, *The Emotional Brain: The Mysterious Underpinnings of Emotional Life* (New York: Simon and Schuster, 1996); and "Emotional Memory Systems in the Brain," *Behavioral Brain Research* 58, 1-2 (1993): 69-79. This material is discussed in Martha Nussbaum, *The New Religious Intolerance: Overcoming the Politics of Fear in an Anxious Age* (Cambridge: Harvard University Press, 2012), pp. 25ff.

²² I surveyed some of these reasons in Chapter Two when discussing epistemic fortification.

of a well-ordered liberal democracy is one designed to serve as our moral lodestar. It is not some prediction about what “the end of history” will entail. It is an articulation of our convictions and the political life they demand. And it is hard to see how our liberal convictions could demand a world in which citizens complacently assumed that they had somehow transcended the epistemic pitfalls that had afflicted all earlier civilisations.

I argued in the last chapter that the citizens of well-ordered liberal democracies are subject to fortificational duties to maintain and strengthen their powers of moral reasoning. In this chapter I explore one such duty: to engage in democratic discourse in a manner that facilitates the improved moral reasoning of all. I do not undertake the defence of this duty; given that citizens have duties to decide what to think on a given issue and vote accordingly, it is not difficult to argue that they ought to undertake this process in whatever manner is most conducive to improving their powers of reasoning and thus arriving at reasonable conclusions. My aim here is the modest but crucial task of inquiring into how such citizens ought to *understand* and *conceive* their shared democratic praxis in order for it to be true that it fortifies their capacity for moral reasoning and enables them to reach more reasonable conclusions about the demands of justice. Now the question of what might give democracy its “epistemic value”²³ is a point of debate among democratic theorists.²⁴ Likewise, the question of democratic *deliberation’s* epistemic

²³ Note that “epistemic value” need not be indexed to any *pre-determined* account of what justice consists in, as we reasonably disagree about what that account precisely entails. Rather, the epistemic value of democracy, for political liberals, is simply its tendency to arrive at *reasonable* policies: those that respect all as free and equal.

²⁴ The most famous proposal is predicated on the proposition that if every member of a group has a greater than 0.5 probability of arriving at the truth on an issue, there will be a direct proportionality between the size of the group and its likelihood of arriving at a correct judgement. The basis of this insight is developed in Marquis de Condorcet, “Essay on the Application of Mathematics to the Theory of Decision-Making,” in *Condorcet: Selected Writings*, edited by Keith Michael Baker (London: Macmillan Publishing Company, 1976). Note that a corollary of this proposition is that if each member of a group has a *less* than 0.5 probability of arriving at the truth on an issue, there will be an *inverse* proportionality between the size of the group and its likelihood of arriving at a correct judgement. Condorcet himself worries that being

value has received considerable scholarly attention.²⁵ These debates have largely been conducted with an eye toward establishing whether democracy is *justified*. I believe that political liberalism already has powerful noninstrumental reasons to favour democracy. What I want to do in this chapter is co-opt the debate in democratic theory over democracy's epistemic value for the purposes of the theory of stability.

Specifically, I shall investigate what I believe to be a powerful but neglected argument in the literature, which David Estlund calls *the democracy/contractualism analogy*: the idea that democracy's epistemic intelligence, if it has it, can be traced to crucial similarities between good democratic politics and the idealised choice situations employed by contractualist liberals to explicate or construct correct principles of justice. On this analogy, the collective political decision-making definitive of democracy is best conceived as an attempt to realise the very process of intersubjective justification that (according to political liberals) defines what is just; therefore, any tendency democracy might have to produce just outcomes could conceivably be attributed to the success of

part of a large group can have deleterious effects on the quality of individuals' reasoning; if he is right, his proposition is an argument *against* democracy's epistemic value. But see the discussions on Condorcet in Robert E. Goodin and Christian List, "Epistemic Democracy: Generalizing the Condorcet Jury Theorem," *Journal of Political Philosophy* 9, 3 (2001): 277-306; Robert E. Goodin and David Estlund, "The Persuasiveness of Democratic Majorities," *Philosophy, Politics & Economics* 3, 2 (2004): 131-42; and Jeremy Waldron, *Law and Disagreement* (New York: Oxford University Press, 1999), pp. 51-53, 134-36. For other arguments on the epistemic value of collective decision-making, see Cass Sunstein, *Infotopia: How Many Minds Produce Knowledge* (New York: Oxford University Press, 2006), and James Surowiecki, *The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business* (New York: Doubleday, 2004).

²⁵ For example, see Joshua Cohen, "Procedure and Substance in Deliberative Democracy," in *Democracy and Difference*, edited by Seyla Benhabib (Princeton: Princeton University Press, 1996), and "Deliberation and Democratic Legitimacy," in *Deliberative Democracy: Essays on Reason and Politics*, edited by James Bohman and William Rehg (Cambridge: MIT Press, 1997); John Dryzek, *Discursive Democracy: Politics, Policy, and Political Science* (Cambridge: Cambridge University Press, 1990), pp. 29-110, and *Deliberative Democracy and Beyond: Liberals, Critics, Contestations* (New York: Oxford University Press, 2000); Robert E. Goodin, *Reflective Democracy* (New York: Oxford University Press, 2003); Amy Gutmann and Dennis Thompson, *Democracy and Disagreement* (Cambridge: Harvard University Press, 1998), and *Why Deliberative Democracy?* (Princeton: Princeton University Press, 2004); Hélène Landemore, *Politics, Collective Intelligence, and the Rule of the Many* (Princeton: Princeton University Press, 2013).

such an attempt. Estlund considers and rejects this analogy. In so doing, he powerfully resists the prevailing inference from *political liberalism* (the idea that correct principles of liberal justice can be construed as the conclusions to a hypothetical choice situation among reasonable citizens) to *deliberative democracy* (the idea that politics should ideally operate along the lines of such choice situation). Therefore, according to Estlund, while liberals should be democrats for many reasons, they should *not* be democrats on the basis of a democracy/contractualism analogy, and they should not be the kinds of *deliberative* democrats suggested by the analogy. I will argue against Estlund, and in so doing, resuscitate the analogy and use it to explain how to think about democracy's epistemically fortifying character.

In Section 2, I ask why contractualist liberals might contend that democracy possesses epistemic value, and I present the proposal that a *democracy/contractualism analogy* (DCA) could justify this contention. In Section 3, I will argue that Estlund's rejection of this proposal relies upon a flawed interpretation of the nature of the particular contractualist situation (T. M. Scanlon's) in reference to which his argument is advanced. By defending an interpretation of Scanlon's contractualist situation according to which it is best conceived as an account of how we ought to understand and engage in first-order moral deliberation, I show how it can ground a partial analogy between the kind of reasoning that transpires among participants in a contractualist situation and the kind of reasoning that citizens should ideally employ in their deliberations with one another as citizens of liberal democracies. In Section 4, I will explain why the conception of participation at work in Scanlon's contractualist situation—in which parties seek to identify mutually justifiable principles, free from prior dogmatic commitments about what justice substantively requires—is precisely the participatory ideal that actual citizens should strive to realise. I close by suggesting that citizens have a fortificational duty to

structure their democratic life in accordance with this ideal as I have described it. By organising democracy in such a way as to improve the reasoning of all, we assuage the epistemic worries that members of the well-ordered society—and, *a fortiori*, that *we*—ought to have about the justification of our convictions.

§2. The Democracy/Contractualism Analogy

David Estlund's *Democratic Authority* is one of the most ambitious recent attempts to answer the questions that have dominated most normative democratic theory since Rousseau: what is the justification for democracy, and why is it the kind of justification that citizens committed enduringly to one another's moral status should accept? Estlund considers numerous approaches to the value of democratic procedures and finds them lacking. If what makes democracy valuable is its procedural fairness—the way it does not stack the deck in favour of anyone's preferred view—why not flip a coin? If what makes democracy valuable is the stage it offers people to express their opinions publicly and therefore have their views taken seriously by their listeners, why bother voting? Estlund thereby adduces pressure to believe that the justification of democracy must at least partially have to do with the moral quality of the outcomes it tends to produce. This argument is especially compelling once we realise that we do not care about democracy's procedural fairness or respect for individual opinions once its procedures are marshalled to produce grave injustice.²⁶

²⁶ Estlund does not consider the justification of democracy I endorsed in Chapter One—according to which democracy secures its noninstrumental value from its role in making just civic relationships possible. I believe this noninstrumental justification rescues democracy from the problems Estlund associates with it, as it accounts for why democracy lacks justification when its procedures are morally misused. My aim is thus not to quarrel with Estlund about the justification of democracy. Rather, it is to co-opt his search for democracy's epistemic value—its tendency to “know” justice—for the purposes of the theory of stability.

In the chapter of Estlund's *Democratic Authority* that concerns me here, his mission is to show why democracy's modest epistemic value—its modest tendency to produce reasonable outcomes—is *not* to be explained by what he calls a *democracy/contractualism analogy* (DCA): the claim that democracy can “produce outcomes that are right by contractualist standards...by promoting the similarity (in certain respects) of actual procedures to the procedure in the hypothetical contractualist situation.”²⁷ While Estlund fails to distinguish between the different versions of a DCA, there are several ways in which to understand what it might entail. The first part of Estlund's discussion mostly addresses the idea of what we can call a *structural analogy*, according to which proper democracy secures justice because the structure of democratic decision-making mirrors the structure of some kind of idealised choice situation (such as Jürgen Habermas's, in which participants have unlimited time to deliberate,²⁸ or Brian Barry's Scanlon-inspired deliberation, in which every participant is equipped with veto power,²⁹ or John Rawls's, in which parties suffer a loss of identity.) This idea, Estlund properly detects, is a dead-end; it would be institutionally infeasible and normatively undesirable to give every actual, breathing citizen a veto over every policy. I will not dwell on this aspect of the discussion due to its evident fruitlessness, and instead move to a different kind of analogy that Estlund discusses, what I call a *reasoning analogy*. On this analogy, what is mirrored in actual politics is not the structure of an idealised situation, but rather the processes of reasoning in which parties to that situation engage.

Whom is Estlund talking about? In his original article on which the relevant chapter is based, he claims to be discussing “a central strand in theories of deliberative

²⁷ Estlund, *Democratic Authority*, p. 239.

²⁸ Habermas, *Moral Consciousness and Communicative Action*, pp. 89ff.

²⁹ Barry, *Justice as Impartiality*, pp. 67ff.

democracy,” though this formulation is not present in the article’s current form.³⁰ The roster he has in mind consists broadly of contractualists, including John Rawls, Jürgen Habermas, Brian Barry, Joshua Cohen, and William Nelson³¹. All these thinkers rely upon some kind of idealised contractualist situation to explicate the content of justice—whether it is the original position (Rawls), the ideal discourse situation (Habermas), the ideal deliberation (Cohen), what Barry calls a “Scanlonian original position” (drawing on Scanlon’s contractualist ethics), or what Nelson refers to as the democratic “tests” a principle must pass—and, Estlund adds, believes that a deliberative democracy modelled in some way *on* that situation is the best way to achieve just outcomes. Because Estlund directs his discussion explicitly at Barry’s Scanlonian version of the democracy/contractualism reasoning analogy, I accordingly restrict my discussion to this particular version of the analogy, which Barry designs to serve equivalent purposes to Rawls’s original position.

The basic insight underpinning the reasoning analogy I defend can be grasped by distinguishing two crucial questions. Firstly, *how do we determine what laws justice demands?* Secondly, *how should we determine what laws we are actually going to implement?* The standard answer political liberals offer to the first question is an account of a *contractualist procedure of reasoning*. While the philosophical elaboration and political implications of the DCA are quite sophisticated, the root idea of the DCA is simple and profound: *given* that the morally proper end of legislation is indeed *justice*, the same procedure of reasoning by which we arrive at conclusions about what justice requires could be part of the way we determine what laws we are actually going to implement. In other words: if we all want

³⁰ Estlund, “The Democracy/Contractualism Analogy,” *Philosophy & Public Affairs* 31, 3 (2003): 387-412, p. 387.

³¹ The only such thinker that has not crossed our path yet is William Nelson, *On Justifying Democracy* (London: Routledge and Kegan Paul, 1980).

the laws that are enacted to be just laws, then let's have the process of reasoning that enables us to identify what justice requires be the same process of reasoning that guides how we decide what to implement; if we do this, there is a hope that our conclusions will secure justice as opposed to injustice, for we will have engaged in the very process of reasoning that leads to the identification of ideally just principles and policies.

This is the basic, rough shape of the best possible argument for a DCA along Scanlonian lines. Note that it does not presume that democracy secures justice by “an invisible hand” (*pace* Estlund³²). Just as a contractualist situation can “go wrong” when philosophers make mistaken inferences, deductions, or judgements in the reasoning process, so too can democracy malfunction when citizens and lawmakers make mistaken inferences, deductions, or judgements in the reasoning process. My claim is a qualified and modest one: if we design our public culture in such a way that the democratic reasoning of citizens resembles the idealised reasoning of parties to an initial contractualist situation, we will render our collective debate more morally intelligent and accordingly make ourselves more likely to arrive at reasonable conclusions about what we ought to do. And we will have done it precisely by reconstructing the process by which one ought to *ascertain* what it is that we have reason to do.³³

³² Estlund, “The Democracy/Contractualism Analogy,” p. 387. The reference is famously from Adam Smith, *A Theory of Moral Sentiments* (Indianapolis: Liberty Classics, 1982), p. 185.

³³ And in so doing, we will also have identified a more plausible candidate for democracy's epistemic value than Estlund's own. Estlund believes that democracy's tendency to effectively avoid *primary bads* (war, famine, genocide, political collapse, etc.) is what confers democracy with modest epistemic value. *Democratic Authority*, p. 167. But this empirical claim—which he explicitly chooses not to defend—is contestable; majoritarian votes have unjustifiably burdened the interests of minorities so frequently in democracy's history that I cannot see how this could be the basis of democracy's modest epistemic value. Perhaps when democracies have successfully and enduringly secured justice, it is because democratic populations were concerned with justice's achievement in precisely the manner that my version of the DCA enjoins.

§3. What Should Citizens Ask Themselves?

I will now examine Estlund's argument against the version of the DCA I seek to defend, a Scanlon-inspired reasoning analogy. I believe that his argument against the reasoning analogy hinges upon an exegetically and philosophically suspect account of the role of the idealised choice situation in Scanlon's contractualism, and that it fails for this reason. I will demonstrate its failure by denying one of Estlund's central claims: that participants to a Scanlonian choice situation must not explicitly ask themselves "the primary question" of what justice requires.

Scanlon's contractualist principle stipulates that "an act is wrong if its performance under the circumstances would be disallowed by any set of principles for the general regulation of behavior that no one could reasonably reject as a basis for informed, unforced general agreement."³⁴ He contrasts his approach with that of Rawls, noting that "the fates of others" become relevant to parties in the original position since the oppressed and marginalised in society may include those for whom the deliberating parties are trustees. But Scanlon emphasises that on his view, the involvement of others in the decision-making is reflected in "a different kind of motivation, namely the aim of finding principles that others, insofar as they too have this aim, could not reasonably reject."³⁵ Their aim is to determine what correct moral principles require through attempting collectively to identify principles acceptable to all.

Note that the idea of an idealised contractualist situation appears only in the earliest announcement of Scanlon's theory, where he refers to the idea of a "choice situation" in which individuals possess "full knowledge of their situations" and "a desire on each of their parts to find principles which none could reasonably reject insofar as

³⁴ Scanlon, *What We Owe to Each Other*, p. 153.

³⁵ *Ibid.*, p. 191.

they too have this desire.”³⁶ Brian Barry develops the idea of such a choice situation into the notion of a “Scanlonian original position,” in which the idea that each participant can reject unreasonably burdensome proposals is conceived as a *veto power* possessed by each.³⁷

Estlund identifies an interpretive problem in explaining how the process of principle-selection in Scanlon’s choice situation is to work. What exactly do the parties to Scanlon’s situation ask themselves? The trick, Estlund contends, is that they *cannot* be construed as asking themselves “the primary question”: the question of “what we owe each other” or, equivalently in Scanlon’s view, the question of what “can be justified to others.”³⁸ For the purposes of this discussion, we can equate this with the question of “what justice requires” and with the question of what principles are “mutually justifiable.”) That is because—Estlund believes—Scanlon’s contractualist situation is intended to play a distinctive role as a *device* that helps *us* as philosophers answer the primary question. But if parties to the situation are simply asking that very question themselves, why do we need them?³⁹ To assign parties the primary question would imperil “contractualism’s distinctiveness.”⁴⁰ This pushes Estlund to wonder whether parties are simply pursuing their own self-interest, rejecting proposals every time one frustrates a personal aspiration. But that cannot be quite right either, for Scanlon explicitly says that the parties “are assumed not merely to be seeking some kind of advantage but also to be moved by the aim of finding principles that others, similarly

³⁶ Scanlon, “Contractualism and Utilitarianism,” p. 127.

³⁷ Barry, *Justice and Impartiality*, 67-72.

³⁸ Scanlon, *What We Owe to Each Other*, p. 7.

³⁹ Note that this is a *different* concern from the idea that Scanlon’s situation is circular. While Estlund conflates these two objections, the latter concerns the substantive role of *judgement* in Scanlon’s theory with respect to how reasons are to be weighed against each other. I shall address this issue later on.

⁴⁰ Estlund, *Democratic Authority*, p. 251.

motivated, could not reject.”⁴¹ Estlund thus sets himself the challenge of conceiving how parties could be “moved” in this way while nevertheless not asking themselves the primary question—a challenge that must be accomplished, he thinks, if Scanlon’s contractualist device is to play any distinctive function from our own first-order moral reasoning.

Estlund outlines a solution with two elements. Firstly—according to what he calls “reasonable self-service”—we should conceive of parties as motivated to reject proposals “only if they themselves have personal reasons against them.”⁴² In other words, “the contractual participants do not reject proposals on the grounds that they are reasonably rejectable by someone or other”—this would be the primary question—“but only for their own personal reasons.”⁴³ Secondly, we should conceive of each party as “prepared not to press a personal reason he has against a proposal if this would leave only alternatives to which others had objections at least as weighty.”⁴⁴ By combining these two elements, Estlund believes that “the idea of reasonable accommodation among the contractors could be brought in without giving them the primary question.”⁴⁵

Estlund’s strategy is promising, and it could work, I believe, if only it weren’t for the following question: *where do the alternatives that the parties consider come from?* This unmentioned question haunts Estlund’s entire discussion. Unless there is some designer dropping proposals into the initial situation, like newly cracked fortune cookies for parties to ponder, then proposals must come from participants. How could this be? Estlund does not say. Perhaps, as suggested by his conception of “reasonable self-service,” parties press their own interests *within reasonable limits*. What does this mean?

⁴¹ Scanlon, *What We Owe to Each Other*, p. 5.

⁴² Estlund, *Democratic Authority*, p. 248.

⁴³ *Ibid.*

⁴⁴ *Ibid.*, p. 251.

⁴⁵ *Ibid.*, p. 252.

Estlund's discussion gives us little guidance; all he says is that contractors accommodate one another purely by deciding not to veto a self-burdening proposal when alternative proposals are even more burdensome to others. We do not know whether he thinks any accommodation would transpire in the phase of *proposing* principles.

When we shift from the matter of rejecting proposals to the matter of advancing them, Estlund's distinction between reasonable self-service, on the one hand, and addressing the primary question, on the other hand, collapses. We have no reason to think that Scanlon's stipulation of parties as "moved by the aim of finding principles that others, similarly motivated, could not reject" applies only to their role in *rejecting* principles; such a presumption would be decidedly *ad hoc*.⁴⁶ Now, when it comes to conceptualising the process of rejection in the contractualist situation, it is perfectly coherent to stipulate that when a participant rejects a proposal, she will only reject it on the basis of personal reasons: if a proposed principle deeply frustrates someone's commitments or interests, only that contractor will be the one who rejects it. Likewise, it is perfectly coherent to say that a participant will take others' interests into account by refraining from vetoing self-burdening proposals that burden others even more. But when we turn to the matter of a participant who *proposes* a principle that advances her interests, we must include the proviso *within reasonable limits* in order for that participant to still be the kind of agent who possesses the aim of finding mutually justifiable principles. And this modifying clause can mean only one thing: that she is advancing her interests by proposing principles that do not unreasonably burden the interests of others. *But to propose principles that advance one's interests within reasonable limits, and to propose principles that no one could reasonably reject, are one and the same.*

Estlund writes:

⁴⁶ Scanlon, *What We Owe to Each Other*, p. 5.

[T]ruly analogous participants will not address the question of justice itself, but only their own interests so far as they can be reasonably pressed. This is not egoism, but nor is it a sufficient orientation to the common good to support the tracking claim under circumstances of real and proper democratic choice...[]Justice would not be directly addressed by participants who were analogous to the hypothetical contractors.⁴⁷

But this can't be right. Pressing one's interests only so far as reasonable means proposing principles that advance one's interests but in a manner that does not unreasonably burden the interests of others. It is not enough to advocate a principle of distributive justice with the aim of advancing one's own economic interests without demonstrating why it could not be reasonably rejected.

With this in mind, the distinction between advancing one's own interests within reasonable bounds and searching for mutually justifiable principles disappears. There is no middle ground between pursuing one's self-interest and pursuing mutually justifiable principles when it comes to the formulation of proposals. Moreover, Estlund's analysis is misleading insofar as it suggests that a participant is *not* advancing her interests when she proposes mutually justifiable principles. After all, to search for mutually justifiable principles—principles that no one could reasonably reject—is to search for principles that are justifiable to *oneself*, as well as to others. Part of explaining that a principle is mutually justifiable means showing why it permits us “to govern ourselves in a way that others could not reasonably refuse to license.”⁴⁸ Principles that would considerably and unnecessarily burden one's own (reasonable) aims are not mutually justifiable. Therefore, to identify mutually justifiable principles just *is* to identify principles that advance one's own interests within reasonable bounds. In this way, Estlund's distinction between “reasonable but self-serving motives”⁴⁹ and the motives to find principles that no one,

⁴⁷ Estlund, *Democratic Authority*, p. 241.

⁴⁸ Scanlon, *What We Owe to Each Other*, p. 157.

⁴⁹ Estlund, *Democratic Authority*, p. 244.

similarly motivated, could reasonably reject, is illusory when it comes to the matter of proposing principles.

Before proceeding to sketch an interpretation of Scanlon according to which we *should* see parties as asking themselves the primary question—and explaining why we should not regard this as unfortunate for the theory—I want to address one way in which Estlund could respond to my criticisms and defend the thesis that parties do not ask themselves the primary question. This would be to suggest that, while parties are accommodative at the rejection stage, we should conceive of parties as advancing principles as *strictly self-interested* when it comes to the stage of proposing principles. This, at first, seems to fit with the rest of Estlund’s account: party A proposes a monstrously self-serving principle that would require the subordination of the rest of humanity to A’s aims; B, burdened by A’s proposal, vetoes it; A politely declines to object in response, recognising the reasonableness of B’s rejection; and the self-serving principle thereby slides into the rejection pile.

One crucial problem with this approach is that it is not clear how we end up determining the set of principles that lie beyond reasonable rejection if all that is happening is that parties propose self-serving accounts with no attention (at the stage of proposals) to reasonableness. The situation would consist in the multiply repeating proposal and rejection of the same candidate principles. So we would need an account of how each party proceeds once her prudentially optimal principles are rejected. What criteria would A use when making his next gambit? Do parties concede territory, as it were, through each iteration of proposals—using information from the previous round of rejections to dilute the self-serving character of each fresh proposal as they bargain their way to justice? If so, how far do they dilute them, and in accordance with what criteria? If the criteria are *moral*, then there is pressure to think that the parties could just

cut to the chase and propose mutually acceptable principles—which would undermine Estlund’s thesis that they are not asking themselves the primary question. I believe that these questions render this proposed interpretation unworkable. But even if they could be answered adequately, this would nevertheless be an extremely exegetically unattractive interpretation of Scanlon’s contractualist situation. Scanlon insists that we should regard parties as “moved by the aim of finding principles that others, similarly motivated, could not reasonably reject.”⁵⁰ It is highly implausible to think that this is somehow an error; Scanlon makes the comment *precisely* when he is contrasting his initial situation with others in the literature: “What distinguishes my view from other accounts involving ideas of agreement is its conception of the motivational basis of this agreement.”⁵¹ Yet the idea that such parties are characterised as *only* prudentially rational by (my extended version of) Estlund’s interpretation of the proposal phase clearly sits in serious tension with this characterisation. It is worth searching for a rival interpretation of Scanlon’s contractualist situation that avoids this tension, and indeed that avoids the serious problems that I have shown to afflict Estlund’s reading.

§4. The Interpretation of Contractualism

4.1. Reinterpreting Scanlon

If my criticism of Estlund is sound, then we should regard Scanlon’s contractors as asking the primary question. This is the only way, I have argued, to make sense of Scanlon’s contention that they are “moved by the aim of finding principles that others, similarly motivated, could not reasonably reject.”⁵² Intriguingly, Estlund himself admits that this quotation “seems to give them the primary question”—though he only admits

⁵⁰ Scanlon, *What We Owe to Each Other*, p. 5.

⁵¹ *Ibid.*

⁵² *Ibid.*

this in a footnote in his 2003 article; it does not appear in *Democratic Authority*. “The exegetical problem raised by these quotes in Scanlon cannot be pursued here,” Estlund continues in that footnote—a puzzling statement, given that the right way to understand Scanlon’s account is precisely the focus of his discussion.⁵³

Recall that what motivates Estlund is a worry that assigning the primary question to contracting parties will end up imperiling “contractualism’s distinctiveness.”⁵⁴ If parties are simply asking the primary question, why do we need to think about *them*? We can just ask ourselves. The choice framework’s role as a device through which conclusions about right and wrong are constructed or determined depends, he thinks, on this not being the case. As a result, Estlund goes to great lengths to interpret Scanlon in a way that retains the distinct character of the reasoning within his contractualist situation. But his efforts, as I have shown, are strained. Moreover, they are unnecessary. For, there is nothing incompatible between the claims (a) that parties to Scanlon’s contractualist situation ask themselves the primary question, and (b) that Scanlon’s contractualist situation represents an account of how correct moral principles are best identified. That is because Scanlon’s contractualist situation just *is* a framework for understanding how moral reasoning—the process of arriving at determinate conclusions about right and wrong—properly proceeds. It is an account of first-order moral reasoning, through which determinate moral principles are identified or, if you prefer, constructed.

To understand this way of interpreting Scanlon, consider some structural similarities between his view and Rawls’s own view. Glossing over details for the sake of simplicity, we can regard Rawls’s project as beginning by identifying (a) an account of citizens’ moral status as free and equal, and (b) a companion account of their interests *as*

⁵³ Estlund, “The Democracy/Contractualism Analogy,” p. 401n.

⁵⁴ Estlund, *Democratic Authority*, p. 251.

free and equal (such as their interests in developing and exercising their moral powers, and thus in primary goods). He then seeks to determine (c) the terms of interaction that respect this understanding of persons as free and equal. The original position is a process of reasoning that takes us from where we start—with (a) and (b)—over to (c). “The leading idea,” Rawls writes in 1980, “is to establish a suitable connection between a particular conception of the person and first principles of justice, by means of a procedure of construction.”⁵⁵ We begin with some account of normative “raw materials,” including an idea of the basic moral status of persons, and we reason our way from that account to specific principles.⁵⁶

Scanlon’s account can be understood similarly. As Scanlon himself says,

The very structure of the [contractualist] test is already a moral principle that constrains the kind of norms that can pass it. It is already a moral principle that everybody counts—that we should be able to justify our norms to everyone.⁵⁷

Scanlon thus begins with an account of the moral status of persons according to which they are equally entitled to justification.⁵⁸ In virtue of this moral status, Scanlon believes—and here I am quoting Michael Ridge—that everyone “must make room for the reasonable agent-relative concerns of others.”⁵⁹ Scanlon thus proceeds to offer us an account of objective but nevertheless agent-relative reasons that each person has—reasons to seek food and shelter, pursue certain projects, devote time to one’s family, develop relationships, etc.⁶⁰ What Scanlon’s contractualist deliberation does is serve as

⁵⁵ Rawls, “Kantian Constructivism in Moral Theory,” p. 304.

⁵⁶ Rawls, *Political Liberalism*, p. 93.

⁵⁷ Alex Voorhoeve and T. M. Scanlon, “The Kingdom of Ends on the Cheap,” in *Conversations on Ethics*, edited by Alex Voorhoeve (New York: Oxford University Press, 2009), p. 182.

⁵⁸ For a promising development of this idea, see Rainer Forst, *The Right to Justification: Elements of a Constructivist Theory of Justice*, translated by Jeffrey Flynn (New York: Columbia University Press, 2011).

⁵⁹ Michael Ridge, “Saving Scanlon: Contractualism and Agent-Relativity,” *The Journal of Political Philosophy* 9, 4 (2001): 472–81, p. 481.

⁶⁰ Scanlon, *What We Owe to Each Other*, p. 204.

the *framework of reasoning* between these normative raw materials from which we begin—an account of the moral status of others and a companion account of persons’ agent-relative reasons—to reach specific conclusions about right and wrong. It describes how we should reason our way from these raw materials to arrive at conclusions about what we owe to each other. And we do this by aiming to devise a set of principles that, given the objective agent-relative reasons for action people have, does not burden anyone in a way that could be avoided by a rival set of principles. We do this, in other words, by *asking the primary question*. Just as Rawls’s original position takes us from abstract normative raw materials to determinate conclusions, so too does Scanlon’s contractualist situation. Each is an account of how someone who endorses the fundamental ideas of the theory—who is morally committed, in Scanlon’s case, to the equal claim of all to justification—should reason her way to conclusions about what she ought to do. Indeed, DCA-pioneer Brian Barry’s influential proposal⁶¹ that we import Scanlon’s contractualist account into Rawls’s theory would only make sense if their accounts of contractualist reasoning played similar roles.⁶²

There are many virtues of interpreting Scanlon in this way. Firstly, it is in many respects a more natural reading. We need not strain to find a description of what goes on “inside” the contractualist situation that marks it out as distinct from what goes on “outside” in the first-order activities of moral deliberation; instead, we can embrace the

⁶¹ Barry, *Justice as Impartiality*, pp. 67-72

⁶² It may appear that Rawls and Scanlon are crucially disanalogous in that Rawls’s parties *cannot* be conceived as asking themselves the primary question. But we can redescribe what goes on inside the original position in a way that eliminates this disanalogy. The original position is a matter of thinking about ourselves as free and rational agents, and then asking what principles secure our fundamental interests *as* free and rational agents, free from the morally arbitrary contingencies that render our choices heteronomous. These principles are the principles of justice. Recall that in the stirring final passage of *A Theory of Justice*, “to see clearly and to act with grace and self-command” from the perspective of the original position (and thereby aspire to “purity of heart”) is precisely what Rawls enjoins us to do (p. 514). Thus when appropriately described, even Rawls’s own procedure of construction can be viewed as one of asking the primary question—and undertaking the reasoning that leads one to settle on answers *to* that question.

idea that Scanlon's descriptions of the ideal agents in *each* are identical. Just a page before Scanlon describes the parties as "moved by the aim of finding principles that others, similarly motivated, could not reasonably reject,"⁶³ he says, remarking on the phenomenology of moral experience:

When I ask myself what reason the fact that an action is wrong gives me not to do, my answer is that such an action would be one that I could not justify to others on grounds I could expect them to accept.⁶⁴

My reading does not force us to invent some way in which the two passages differ. The idealised processes of reasoning that carry on inside the contractualist deliberation are the same processes of reasoning that we ourselves undertake when we do moral deliberation properly.

Secondly, my interpretation helps to make sense of Scanlon's mysterious contention that reasons of "fairness" can be invoked against principles in the contractualist situation. This is because principles that "arbitrarily favor one person over others" offend against the root idea of the equal stranding of all persons that serves as the foundation of Scanlon's contractualism.⁶⁵

Thirdly, this interpretation helps immunise Scanlon from the common objection that his theory is objectionably circular because it relies on substantive judgements. This common objection begins by recognising that the reasonableness of any given rejection in Scanlon's deliberation needs to be assessed somehow—namely, by assigning weights to agent-relative reasons people have for and against a proposed principle, and seeing which person's reasons are strong enough to win out. But if this is true—and it is—a considerable amount of important work thereby transpires at the level of assigning moral weights to the different reasons, assignments which cannot be explained by the final

⁶³ Scanlon, *What We Owe to Each Other*, p. 5.

⁶⁴ *Ibid.*, p. 4.

⁶⁵ *Ibid.*, p. 212.

principles of morality that are settled upon since they are necessary in *reaching* those final principles. What this objection concludes is that this makes the idea of a contractualist situation entirely dispensable: all we need to do is look at the realm of reasons for action, judge the relative strengths, and then see what principles are demanded or disallowed by the strongest reasons.⁶⁶

However, once we see that Scanlon's account does not cast the contractualist situation as a specialised device, but rather a *framework* for understanding moral deliberation at its best—a framework that directs and orients these judgements in reasoning—this objection misses the mark. Scanlon is explicit about the role of *judgment* in his theory:

According to my version of contractualism, deciding whether an action is right or wrong requires a substantive judgment on our part about whether certain objections to possible moral principles would be reasonable⁶⁷

But this is not circular; it simply comes with the territory of all accounts of first-order moral reasoning (utilitarian, virtue-ethical, and Kantian alike). Rawls's account, too, depends on judgement in determining how to lexically rank different principles. That there are such burdens of judgement in any theory is precisely why there is reasonable disagreement about what conclusions are properly reached—but this is an issue for *any* view.⁶⁸ Once Scanlon's account is understood as a description of idealised moral reasoning, the objection that such reasoning depends on substantive judgements becomes utterly unsurprising, and thus loses its force.⁶⁹

⁶⁶ Nicholas Southwood, *Contractualism and the Foundations of Morality* (New York: Oxford University Press, 2010), pp. 61-69.

⁶⁷ Scanlon, *What We Owe to Each Other*, p. 194.

⁶⁸ Rawls *Political Liberalism*, pp. 56-57.

⁶⁹ It is true that Scanlon is concerned to be sure that his framework is not “unnecessary” (*What We Owe to Each Other*, p. 213). But here he is concerned not with the role of judgment, but with the description of moral motivation: we must not describe idealised reasoners in such a way that they reason with substantively developed views of justice *already* on hand. That *would* make the

Fourthly, the most important reason to think that Scanlon would endorse this interpretation is that he *does* endorse it. Consider a restated form of the aforementioned circularity objection posed by Alex Voorhoeve:

It seems that to get some determinate answers to questions of right and wrong, we must first decide which things count as valid ground for rejecting a principle. But this would appear to leave the procedure devoid of any content: any moral principle you pull out of the contractualist hat is put there at the outset when you decide which things count as grounds for rejecting a principle.⁷⁰

In response to this, Scanlon replies—and this is the kicker:

[W]e can't avoid questions of judgement, and this could lead to the charge that the contractualist reaches conclusions only because he has helped himself to lots of substantive ideas along the way, and that the theory itself therefore doesn't really yield new answer. *It may be that this criticism is justified.* There are different desiderata of a moral theory. I am inclined to think that providing a way of cranking out novel principles is overrated [...] *What I mean to do is offer a way of understanding what moral thinking is.*⁷¹

Scanlon goes even further, disputing the idea that parties to his contractualist situation are best conceived (as Estlund recurrently does) as a “hypothetical agreement” that therefore must be modelled in some special way distinctly from our normal moral deliberation. Instead, Scanlon argues:

To describe [my theory] as a ‘hypothetical agreement’ is already mistaken. It isn't about what people would agree to under certain conditions. Rather, what matters is what it would be *reasonable* for them to reject under certain conditions, that is, if they, too, were trying to find principles that others could not reasonably reject.⁷²

That such conditions are hypothetical therefore only serves to indicate that they are idealised—they characterise the conditions of faultless reasoning—not that they refer to

framework unnecessary. But this proves my point: for moral deliberation would *also* be “unnecessary” if we had already decided and had made up our minds unreflectively about what we owe to each other. Scanlon's parties, like us, should be armed with two things only: a capacity to make substantive judgements about reasons, and the willingness to search for mutually acceptable principles. These are a far cry from having a determinate account of justice already on hand. We still need a reasoning framework to get us there.

⁷⁰ Voorhoeve and Scanlon, “The Kingdom of Ends,” p. 182.

⁷¹ *Ibid.*, p. 183.

⁷² *Ibid.*, p. 184.

the operations of mysterious, idiosyncratically motivated agents who undertake some special contract with one another. Rather, they refer to *our* moral reasoning about what we owe to each other, at its best.⁷³

Pessimistically, Michael Otsuka writes that Scanlon's contractualist situation "is bought at the price of a reduction in the usefulness of the contractualist device as a means (distinct from ordinary forms of moral reasoning) of arriving at the right principles of justice."⁷⁴ But Otsuka here makes the same mistake of Estlund: assuming that Scanlon's contractualist situation was a specialised device to serve as a "means" to philosophers external to our first-order moral reasoning. On the contrary: it is an account of how we ought to *undertake* first-moral order moral reasoning, one rooted in Scanlon's thesis that "the idea of justifiability to others is taken to be basic" to what morality *is*, and thus how we should undertake the search to identify its demands.⁷⁵

4.2. From the idea to the practice of reasonable rejection

Having presented an interpretation of Scanlon's contractualism that regards it as a description of how we ought to understand and engage in first-order moral deliberation, I want to note some implications of this interpretation that are relevant for the defence of a Scanlon-inspired DCA. Scanlon's discussion of deliberation is clearly focused on the idea of deliberation within a single person's mind: each asks herself whether it could be reasonable for someone to reject the principle upon which she is about to act, and draws

⁷³ Aaron James, who casts Scanlon as a constructivist alongside Rawls, Christine Korsgaard, and Onora O'Neill, writes that Scanlon's contractualism is "concerned to *characterize central features of moral reasoning* [...] which appear across a wide variety of cases, in a way which nevertheless might guide judgment about principles in specific contexts." This supports my conclusion. Aaron James, "Constructivism, Moral," in *The International Encyclopedia of Ethics*, edited by Hugh LaFollette (Oxford: Blackwell, forthcoming), p. 6.

⁷⁴ Michael Otsuka, *Libertarianism Without Inequality* (New York: Oxford University Press, 2003), pp. 5-6.

⁷⁵ Scanlon, *What We Owe to Each Other*, p. 5.

a personal judgement about its justifiability. But while Scanlon is right that such judgments are ones that “each of us must make for him- or herself,” he also emphasises that “interaction with others plays a crucial role in arriving at well-founded moral opinions.”⁷⁶ What I want to do here is show how Scanlon’s account of moral deliberation presents a useful organising framework for those interactions in actual political life, one that enables us to see how we might best operationalise our quest for epistemic fortification.

Scanlon enjoins us to search for principles that no one could reasonably reject. But the *could* is crucial. After all, if participants to the moral deliberation really are the kinds of participants who seek mutually acceptable principles, why would there need to be any reasonable rejections that transpire at all? Who in the deliberation would have proposed a principle that was reasonably rejectable, and why would she have proposed it if she had the aim of finding mutually justifiable principles? If everyone reasons faultlessly in moral deliberation, and if we assume that all the addressed questions have determinate answers, then this will not happen. But, of course, real deliberations *will* include moral mistakes. And when such mistakes are made, the right response in deliberation is to condemn them on the basis of reasons—to “reasonably reject” them. The powerful image of people who stand and actively *reject* proposals on the basis of their unreasonableness is indispensable to the democratic recreation of Scanlon’s process of reasoning. Such active rejections serve two vital roles in deliberation. The first is to expose malevolence: that people are wilfully acting by norms they cannot justify to others. But the second—what I want to pursue here—is epistemic. When we ask ourselves *why* someone in a real-life deliberation who is morally motivated might have

⁷⁶ *Ibid.*, pp. 393-94.

proposed a principle that was reasonably rejectable, the answer is clear: because she did not know it was unreasonable.

I want to develop the proposal that Scanlon's contractualism provides not just a framework for individual moral deliberation, but constitutes a powerful proposal for *collective* moral deliberation. Let us quickly examine how this framework operates in greater detail. When a proposal is advanced, he claims, the first task is to identify whether any participant has a justifiable objection to the proposed principle. If a participant does—if the proposal burdens her in a way that alternative arrangements could avoid without burdening another participant more—the proposal must be rejected. In making such a determination, it must also be considered whether the *failure* to permit this principle would have far worse effects on others; in such a case, those others might respond with *objections to the proposed prohibition* of the proposed principle.⁷⁷

This idea of a dialectical exchange of considerations that results in a judgement about whether a proposed rejection of a proposed principle is reasonable is also reflected in Scanlon's discussion of *generic reasons*, or reasons that are grounded in “commonly available information about what people have reason to want...in virtue of their situation, characterized in general terms.”⁷⁸ It is on the basis of these generic reasons that the proposed rejections of principles are justified—are deemed *reasonable* rejections. Consider how the process of generalisation transpires. Participant A advances a proposal whose adoption would advance her interest, but believing that it does not unreasonably burden others. Participant B then contends that A's proposal does, in fact, burden B's interests in a manner that an alternative arrangement would not; B then proposes this

⁷⁷ *Ibid.*, pp. 202-203.

⁷⁸ *Ibid.*, p. 171. Among such generic reasons, Scanlon argues, are “reasons to want to avoid bodily injury,” “to be able to rely on assurances they are given,” “to have control over what happens to their own bodies,” and so on, including being able to “give special attention to our own projects, friends, and family” (p. 204).

alternative arrangement. Participant C then objects that the new arrangement is overly burdensome to her, and that a further alternative is available.

Now fill in the example with the following case (albeit it is not one that would obtain in the a well-ordered society). Perhaps A proposed that her Catholicism be the religion of the state, thinking that it would allow to practice her Catholic beliefs easily, and being unaware that there are others who would object strongly; B, a Protestant, responds by suggesting that they generalise the proposal to make Christianity the official religion of the state, so as not to burden B's interests in practicing her own sect's faith; and then suppose C, a Buddhist, suggests they generalise the principle further to be a principle of religious freedom, whereby all can exercise their religion of choice. They have now arrived at a principle that no one could reasonably reject, and specifically through the identification of the *generic reasons*.

The upshot of the idea of generic reasons is that an agent who rejects a proposed principle must explain the burden that the acceptance of such a principle would inflict upon her in terms all could coherently relate to. For example, if a person felt compelled to consume a particular kind of food in accordance with her religious beliefs, and a principle was proposed that would have the consequence of banning the consumption of that food, she would have to explain her plight in mutually comprehensible terms by claiming that a principle forcing someone to sacrifice a core practice endorsed by that person's conscience, in the absence of further evidence that the practice's exercise burdens others in any comparable way, is a reasonably rejectable principle. But in order to endorse her situation as such, I maintain, one would have to roughly comprehend the role of the particular food in her religion, and the fact that she actually endorsed it as central to her core beliefs, in order to understand her argument, or to confirm that her particular situation would indeed be protected by the generic value to which she is

appealing. Cases of this kind abound from political experience; the justifiability of serving pork in state school cafeterias would never have been considered an issue of justice before some began to question it. In such cases, each side must perform important mental work: the objector must do the work of re-describing her situation under a generic reason, like dietary autonomy or liberty of conscience; and everyone else must realise that the practice in question is correctly protected by the generic reason—even if the practice is one predicated on specific values they do not share, or even that they potentially repudiate in their own lives.

Indeed, it is precisely because the particular claims subsumed under generic reasons are often subject to intense *epistemically reasonable disagreement* that the process of working out what kinds of claims are or are not covered by generic reasons is particularly onerous. Consider Scanlon's example of a generic reason: *privacy*. People have many different reasons for favouring privacy, but their underlying normative accounts of its importance vary considerably. Therefore, someone who objects to a situation because it would subject her to shame and humiliation—a conservative Muslim woman in a Western airport forced to undergo a strip-search and thereby expose bare skin to passing men, for example—needs to redescribe her needs in mutually comprehensible terms, in generic values. This redescription will be especially important since many Westerners do not object to such strip-searches on privacy grounds. They are thus less inclined to see the objection as covered by the generic reason of privacy, and precisely since so many of them disagree reasonably (in the epistemic sense) with the underlying principle that women ought not expose bare skin to passing men. But the redescription is motivated by her aim to find mutually justifiable principles: to find principles that protect her interests within reasonable bounds and are defended in a way that demonstrates their reasonableness by being cast in a general language that all in the moral community can

share. Only through this process can we come to see whether a particular policy respects persons as free and equal, given the burdens of judgement, or not.

This constructive discussion of the “dynamics” of Scanlonian reasonable rejection helps us construe the practical instantiation of Scanlon’s moral deliberation as a search for *moral learning among participants*. I say that this is “constructive” because Scanlon does not devote much space to the practical development of moral commitments on the basis of reasoning. But he clearly recognises its importance:

[T]he ‘shaping role of the aim of justifiability to others is a dynamic one. There is no fixed list of ‘morally relevant considerations’ or of reasons that are ‘morally excluded.’ The aim of justifiability to others moves us to work out a system of justification that meets its demands, and this leads to a continuing process of revising and refining our conception of the reasons that are relevant and those that are morally excluded in certain contexts...[It] is a continuing process, not a fixed list of results.⁷⁹

Thus Scanlon’s contractualist situation constitutes a promising framework for how morally motivated actors can expose one another’s moral mistakes and learn from one another in actual political life.

It may seem strange to spend so much energy in a chapter on democracy in interpreting the work of T. M. Scanlon, whose work is not addressed to democratic theorists. But aside from the distinctive value of Scanlon’s work to the broader contractualist project of political liberalism, Estlund has made this focus necessary. Estlund has, in many eyes, hammered a nail in the coffin of the democracy/contractualism analogy. But this is mistaken: the nail did not reach its target, and the contents of the coffin remain alive. Debunking Estlund’s flawed rejection of the DCA is the key to resuscitating it as a powerful candidate explanation of how democracy could serve to bring the epistemic best out of us all.

⁷⁹ *Ibid.*, pp. 157-58.

Conclusion

The basis of an analogy between the contractualist situation—in which idealised moral reasoners address the primary question—and actual democratic politics—in which *we* address the primary question, endeavouring to emulate that open-minded search for public justification—should by now be considerably clearer. I have accomplished this chapter's aim: to shed some light on how the fortificational duty to pursue democratic debate in a manner organised to improve the reasoning of all is best operationalised.

By explaining how the *deliberative* character of deliberative democracy is properly explained by concerns of epistemic fortification—because such a democracy so organised makes us into the kind of people who reason well about justice—I have proposed to build the theory of stability into the bedrock of liberal democracy's self-understanding. But we cannot rest on our laurels. Even conscientious practices of citizenship can sometimes go astray, resulting in legislative proposals that threaten to violate individual rights. How can we design the *institutional context* in which citizens and officials deliberate to militate against large mistakes of this kind? It is to this question that I now turn.

CHAPTER FOUR

Democratic Stability and Judicial Review

“It is by no means contrary to practice for laws to be so firmly established that even the king himself cannot repeal them. Even kings have followed the example of Ulysses; they usually instruct their judges to have no respect for persons in administering justice, not even the king himself, if by some odd mischance he commands something which they know to contravene established laws. For kings are not gods, but men, who are often enchanted by the Siren’s song. Accordingly, if everything depended on the inconstant will of one man, nothing would be stable.”¹

– Spinoza

The practices of democratic engagement I discussed in the previous chapter ought to be fixtures of a well-ordered liberal democracy comprising morally motivated but fallible citizens: citizens who realise that their capacities for proper moral reasoning and decision-making are susceptible to attack from manifold countervailing forces. By showing how democracy itself is partly justified on fortificational grounds, I have proposed to build the theory of stability into the bedrock of liberal democracy’s self-understanding. But if citizens are sufficiently fallible that the constant invitation to justify their convictions and conduct to others is necessary for them to sustain their commitment to justice, how can we be sure that even a pervasive ethos of democratic contestation will be sufficient in the well-ordered society to make justice endure? Spinoza worries of the caprice of but one man. Unless we are inclined to think that normal citizens feature fiercer moral fortitude than Ulysses, we should wonder whether democracy—at its worst—could serve to *multiply*, rather than mitigate, our fallibility. There is, after all, only so much evil one human can do. But together, our capacity for injustice is infinite.

¹ Cited in Jon Elster, *Ulysses Unbound* (New York: Cambridge University Press, 2000), p. 89. A less attractive translation can be found in Benedict de Spinoza, *The Chief Works of Benedict de Spinoza*, revised edition, translated by R. H. M. Elwes, (London: George Bell and Sons, 1891).

This chapter examines what to do when democracy morally malfunctions. Specifically, it examines the fortificational duties we have to *prepare* for that eventuality through designing and sustaining institutional mechanisms that keep our shared democratic reasoning on track. Focusing on the practice of rights-based judicial review of legislation as an instructive case study, the chapter defends judicial review as an institution that democratic citizens and their officials have reason to create, or otherwise maintain, for the purpose of disciplining democratic reasoning in the face of citizens' and legislators' enduring susceptibility to moral failure. Judicial review is thereby conceived not as opposed to democracy, but as a didactic institutional crutch that helps to keep democracy oriented toward the achievement of justice. However, as I shall argue, it can only serve this role—without falling prey to the most compelling scholarly objections against it—when the injustice of the legislation it condemns is beyond the pale of reasonable dispute. When judicial review transcends this limited role, it affronts our self-understanding as free and equal citizens.

The structure of the chapter is as follows. In Section 1, I offer some preliminary comments. In Sections 2 and 3, I review the state of play within the contemporary debate on this topic, outlining the vision of *expansive judicial review* defended by Ronald Dworkin and presenting the contrasting view—*judicial review abolitionism*—advanced by Jeremy Waldron. In Section 4, I present a preliminary sketch of my intermediary position. In Section 5, I consider an objection to my account—that it affronts democracy's intrinsic value. By examining what is both right and wrong about this objection, I proceed to offer an institutional proposal that accommodates it in Section 6.

§1. The Practice of Judicial Review

I begin by outlining my subject and issuing some important preliminary comments. Firstly, the topic of concern here is not judicial review of executive or administrative action (itself relatively uncontroversial among political philosophers), but *rights-based judicial review of democratic legislation*: the legal practice according to which “appointed judges have the power to declare enactments of legislatures and other representative institutions void because they offend constitutional guarantees of individual rights.”²

Secondly, the *strong* form of judicial review that judges possess in these jurisdictions is to be contrasted with a *weak* form of judicial review, whereby judges may flag perceived inconsistencies between democratically enacted legislation and fundamental rights without any legal authority to strike down the decision (as with the UK Supreme Court’s power to issue “declarations of incompatibility”³). The scholarly debate on judicial review has all but exclusively concentrated upon the practice’s strong variety. This can be instructive: if an argument for strong judicial review succeeds, then *a fortiori* the case for weak judicial review is secure. But one effect of this focus has been a paucity of reflection on how the strong and weak variants of judicial review differentially impact the moral quality of citizens’ democratic relationships. So although I, too, begin this chapter by discussing the strong case of judicial review—engaging the debate on its usual terms—I shall broaden the discussion later to involve weak judicial review.

² Ronald Dworkin, “Rawls and the Law,” in *Justice in Robes* (Cambridge: Harvard University Press, 2006), p. 255. How we ought legally to understand what such voiding consists in is a matter of controversy. Whether it is a matter of “declining to apply a statute in a particular case”, “to modify the effect of a statute to make it conform with individual rights”, or “to actually strike a piece of legislation out of the statute book altogether”, is clear in some jurisdictions and unclear in others. Jeremy Waldron, “The Core of the Case Against Judicial Review,” *The Yale Law Journal* 115, 6 (2006): 1346-1406, pp. 1354ff.

³ Human Rights Act, 1998, c. 42, § 4(2), (6).

Thirdly, normative political theorists often talk of judicial review as though it were a straightforward process whereby judges look at a statute and ask themselves whether justice permits or forbids it. Corey Brettschneider invites us to adopt this picture by imagining two justices—Justice Process and Justice Results—who decide whether to uphold a given law either on the simple fact of whether it was enacted democratically (Justice Process) or whether it is permitted or forbidden by justice (Justice Results).⁴ But the actual task of constitutional interpretation in real democracies is vastly more complicated. Here I make two assumptions: firstly, I assume that judges ought to make decisions that cohere (in some reasonable understanding of that phrase) with precedent; and secondly, I assume that judges ought sometimes to engage in judgements of moral and political principle when declaring the meaning of constitutional provisions that include abstract moral language. The arguments to come will be oriented around *how* judges should engage in such processes of abstract moral analysis. “Lawyers”—at least when it comes to such analysis—“are always philosophers.”⁵

Finally, the argument to be advanced here takes a stand on controversial issues in normative political theory. But it seeks to avoid adopting any controversial stands in issues in analytical jurisprudence about the necessary and sufficient truth conditions for propositions of law. That I embrace Dworkin’s contention that judges ought to be engaged in the interpretation of the abstract moral provisions embedded in the law does not necessarily commit me to his controversial doctrine of law as integrity.⁶ Legal positivism is wholly consistent with everything I shall say.

⁴ Corey Brettschneider, “Balancing Procedures and Outcomes Within Democratic Theory: Core Values and Judicial Review,” *Political Studies* 53, 2 (2005): 423–42, p. 425ff.

⁵ Dworkin, *Law’s Empire* (Oxford: Hart Publishing, 1998), p. 380.

⁶ *Ibid.*

§2. Expansive Judicial Review

Ronald Dworkin's defence of judicial review flows from an ideal of a self-governing political association in which all are treated with equal respect and concern.⁷ He notes that this ideal is ambiguous between various forms of constitutional design. Addressing how to adjudicate between rival frameworks, Dworkin writes:

I see no alternative but to use a result-driven rather than a procedure driven standard for deciding them. The best institutional structure is the one best calculated to produce the best answers to the essentially moral question of what [the demands of political morality⁸] actually are, *and to secure stable compliance with those conditions.*⁹

Dworkin concedes fully that citizens “can be expected to disagree about which structure is overall best,” and at times he suggests that his allegiance to judicial review in the United States flows from the simple fact that such institution was “distributed by history.”¹⁰ But the thrust of his writing on the topic is clear. “The United States,” he writes, “is a more just society than it would have been had its constitutional rights been left to the conscience of majoritarian institutions.”¹¹

What makes Dworkin's position on judicial review distinctive is his answer to the following question: *what must judges take themselves to be doing in their judicial reasoning for it to be true that their review of legislation will tend toward a more successful achievement of substantive liberal-egalitarian justice than were they not engaging in such review?* Dworkin's answer to that question comes in three parts. (As Dworkin focuses his attention on U.S. constitutional

⁷ Dworkin, *Sovereign Virtue: The Theory and Practice of Equality* (Cambridge: Harvard University Press, 2000), p. 1.

⁸ Dworkin does not use the term “political morality” but instead “the democratic conditions”—the conditions that characterise a self-governing political community in which the state treats its citizens with equal concern and respect. I omit his wording here for expositional ease, since at this early stage in the analysis it could cause confusion for readers unfamiliar with Dworkin's definition of “the democratic conditions.” I discuss his understanding of these conditions below.

⁹ Dworkin, *Freedom's Law* (Cambridge: Harvard University Press, 1996), p. 34. Emphasis added.

¹⁰ *Ibid.*, pp. 34–35. Despite his avowed commitment to interpreting institutional structures handed down by history, Dworkin endorsed the introduction of judicial review in the UK in Ronald Dworkin, *A Bill of Rights for Britain* (London: Chatto & Windus, 1990).

¹¹ Dworkin, *Law's Empire*, p. 356.

interpretation, we shall, as well.) Firstly, the U.S. Constitution contains substantively correct fundamental ideas; we know Dworkin thinks this because he argues his own ideal of justice is partly embedded within it.¹² Secondly, judges ought to interpret the abstract moral provisions of the U.S. Constitution in accordance with the best understanding of that underlying ideal's demands—not assessed in the abstract, but assessed within the network of principles and precedents that comprise the polity's legal tradition. Judges ought to look *backwards* to enable a sense of coherence within the tradition and to treat plaintiffs and defendants fairly. But thirdly, judges must also look *forwards* toward justice, conceiving themselves as part of “an unfolding political narrative.”¹³ Thus when a judge confronts a decision as to whether a particular law is constitutional, she should ask herself what possible answers to that question rest in harmony with the legal tradition, and she should proceed to select the candidate answer that best serves substantive liberal-egalitarian justice.

Dworkin's view is thus clearly different from how political philosophers often caricature the judge who engages in judicial review. Dworkin's judge is not asking herself what the best conception of justice demands, and then discovering such demands in the Constitution. Indeed, Dworkin explicitly states that judges may not, for example, interpret an economic right to distributive justice out of thin air. To do so would strain the demands of *integrity*, which enjoins a principled reconciliation between the fairness of adhering to past tradition and the demands of justice.¹⁴ This is not to say that judges' convictions have no role. Dworkin himself believes that judges are doing their job when they identify a textually unmentioned right to privacy as embedded in the Constitution

¹² Dworkin, *Freedom's Law*, p. 8.

¹³ Dworkin, *Law's Empire*, p. 225.

¹⁴ Dworkin, *Freedom's Law*, pp. 10, 36.

and use it to ground an individual right to abortion.¹⁵ Conceiving precedent as judges' anchor, on the one hand, and justice as the wind that directs their sail, on the other, Dworkin urges: "we must not exaggerate the drag of that anchor."¹⁶ For, Dworkin insists, if the Constitution is committed to the ends of liberal justice (as he believes it is), "it must respect whatever freedoms are indispensable to those ends, *including but not limited to the freedoms more specifically designated in the document.*"¹⁷ Thus it is clear why Dworkin contends that proper judicial review advances substantive liberal justice: because that's why judges ought to think of themselves as aiming to achieve.

Before moving on to see how Dworkin's argument is susceptible to a serious criticism, note how he understands the relation between judicial review and *democracy*. By stipulating that judicial review is an appropriate aspect of political institutions, Dworkin does not take himself to be compromising his account's democratic credentials. A "constitution of principle," Dworkin writes, "enforced by independent judges, is not undemocratic."¹⁸ "[D]eference to temporary majorities on matters of individual right is," Dworkin contends, "brutal and alien."¹⁹ This inflationary approach to democracy's definition has been followed by Rawls,²⁰ Habermas, and many others. So obvious is this point among judicial review's defenders that Samuel Freeman has said anyone who thinks judicial review *is* antidemocratic is being "disingenuous".²¹ In what follows, I will

¹⁵ *Ibid.*, p. 112ff.

¹⁶ *Ibid.*, p. 11.

¹⁷ *Ibid.*, p. 8. Emphasis added.

¹⁸ Dworkin, *Law's Empire*, p. 123.

¹⁹ *Ibid.*, p. 71.

²⁰ "The form and content of [public] reason...are part of the idea of democracy itself." Rawls, "The Idea of Public Reason Revisited," pp. 440-41.

²¹ Samuel Freeman, "Constitutional Democracy and the Legitimacy of Judicial Review," *Law and Philosophy* 9, 4 (1990-91): 327-70, p. 367.

not pursue this “definitional stop,”²² which terminates the debate over judicial review’s tension with democracy by nomenclatural fiat. We must see if there is anything to be said for critics’ protestations, to which I now turn.

§3. Judicial Review Abolitionism

Jeremy Waldron joins Alexander Bickel in proclaiming judicial review “a deviant institution in the American democracy.”²³ While Waldron has advanced many distinct objections against judicial review, the most influential and important among these objections begins by noticing the fact of *reasonable disagreement* about justice.²⁴ Because citizens reasonably disagree, any act of judicial review is not likely to be hailed as a heroic rescue of citizens from their own moral folly, but an unwarranted replacement of citizens’ considered democratic convictions about justice with the imposed convictions of a judicial “epistocracy.”²⁵ In permitting such replacements, judicial review “disenfranchises ordinary citizens and brushes aside cherished principles of representation and political equality in the final resolution of issues about rights.”²⁶ So described, Waldron’s argument against Dworkin is simple and arresting. Especially once we see that judges on the court are having the exact same debates about justice as the citizens are—and then settle their disputes by “head-counting”²⁷—judicial review appears indefensible indeed.

²² H. L. A. Hart, “Prolegomenon to the Principles of Punishment,” in *Punishment and Responsibility: Essays in the Philosophy of Law*, second edition, edited by John Gardner (New York: Oxford University Press, 2008), p. 5.

²³ Alexander Bickel, *The Least Dangerous Branch: The Supreme Court at the Bar of Politics*, second edition (Binghamton: Vail-Ballou Press, 1986). pp. 16-17.

²⁴ Waldron, “The Core of the Case,” p. 1368.

²⁵ Estlund, *Democratic Authority*, p. 206.

²⁶ Waldron, “The Core of the Case,” p. 1353.

²⁷ *Ibid.*, p. 1391.

To pinpoint what is right—and what is wrong—with Waldron’s objection to judicial review, we must understand how his objection sits within his complex democratic theory. At first glance Waldron appears to value democracy intrinsically. He argues that it respects citizens by granting each equal say in determining what ought to happen in the face of disagreement, “pay[ing] tribute to their fundamental equality”²⁸—not merely by affirming the equal moral worth of every person, but by honouring their equally authoritative capacities to reason thoughtfully about what it even *means* to respect persons’ equal worth.²⁹ That this is not exhaustive of Waldron’s position is clear from his agreement with Jon Elster that purely “expressivist” justifications of democracy are inadequate, privileging “agonistic flair” and “catharsis” over “the substance of politics—the decisions to be made and their implications for real people.”³⁰ Democracy does not exist for the *purpose* of honouring citizens’ equal status; it is no civic ritual. It exists to make decisions in the face of disagreement about what we should do. Waldron calls this *the circumstances of politics*:

[T]he felt need among the members of a certain group for a common framework or decision or course of action on some matter, even in the face of disagreement what that framework, decision or action should be...³¹

The distinctive appeal of *democracy*, then, is that it enables some decision to be made in a manner that is compatible with honouring the equal moral status of even those who disagree with the ultimate decision. Waldron values democracy because of its capacity to solve important problems in ways that respect the people whose problems they are—a

²⁸ Waldron, *Law and Disagreement*, p. 249.

²⁹ *Ibid.*, p. 251.

³⁰ Waldron, *Law and Disagreement*. Here Waldron is following Elster, *Sour Grapes: Studies of in the Subversion of Rationality* (Cambridge: Cambridge University Press, 1983), p. 99. For the expressivist view on democracy, see Stanley Benn, “The Problematic Rationality of Political Participation,” in Benn, et al., *Political Participation* (Canberra: Australian National University Press, 1978), p. 19.

³¹ Waldron, *Law and Disagreement*, p. 102.

view I find profoundly attractive, given the similarity to Eric Beerbohm's position, which I embraced in Chapter One.

"Legislation deserves respect," Waldron writes, "because of the achievement it represents in the circumstances of politics: action-in-concert in the face of disagreement."³² Judicial review, he thinks, short-circuits that achievement by supplanting the controversial verdicts of a respectful democratic process with the controversial verdicts of a disrespectful anti-democratic process.³³ This conclusion, however, depends on whether Waldron is correct to suppose that democratic legislation really is the sort of achievement worthy of veneration. Waldron says virtually nothing about the moral quality of the decisions democracy makes; he speaks simply of their "moral urgency and importance".³⁴ This is partly unsurprising, given his emphasis on disagreement. But Waldron's terminology invites questions that his account cannot satisfactorily answer: is *every* piece of legislation deserving of respect? Is *every* piece of legislation an achievement?

The ambiguity partly stems from a failure to specify what, exactly, the character of the "felt need" among citizens actually is. It surely is not a *descriptive* understanding of that term; Waldron is not merely reporting that people in democracies just happen to think they need some solution. It is *normative*: "We need settlement not so much to dispose of the issue—nothing can do that—but to provide a basis for common action when action is necessary."³⁵ But what counts as necessary? Waldron argues that whereas "we do not need society-wide settlement" on "the meaning of Hamlet" or "the value of a purely contemplative life," "we *do* need settlement" on "issues of rights."³⁶ But why? The answer can only be that we are morally required to *find* some provisional solution to our

³² *Ibid.*, p. 108.

³³ Waldron, "The Core of the Case," p. 1375.

³⁴ Waldron, *Law and Disagreement*, p. 117.

³⁵ Waldron, "The Core of the Case," p. 1369.

³⁶ *Ibid.*, p. 1370. Emphasis added.

disputes about rights, whereas we are not morally required to affirm (indeed, are disallowed from affirming in the name of all) a common answer to such contestable questions of the good. Thus—while not wishing to deny that we ought to have a “felt need” to bargain with madmen in some deeply non-ideal *modus vivendi*—the “felt need” that Waldron describes is best understood to be the *morally enjoined* felt need experienced by reasonable members³⁷ of the well-ordered society for some common solution to their reasonable disputes.

The problem, as we saw in Chapter One, is that reasonable citizens do not afford respect to *any* old piece of legislation that provisionally settles some disagreement about justice. Rather, they will afford respect only to what they regard as *acceptable* solutions to their disagreements. Coordinated solutions are only valuable when they resolve disagreements in a way that leads to *reasonable outcomes*: outcomes that are plausible interpretations of the ideal of treating citizens as free and equal.³⁸ Waldron notes that the need to determine what ought to be done when the goodness of outcomes is disputed is one of the major contributions of Hobbes to political philosophy, who helped us realise that “any theory that makes authority depend on the goodness of political outcomes is self-defeating, for it is precisely because people *disagree* about the goodness of outcomes that they need to set up and recognize an authority.”³⁹ But even *Hobbes* believes that human beings’ obligations to obey the legitimate sovereign disappear (indeed, the legitimate sovereign ceases to *be*) the moment the sovereign stops protecting them from violent death:

³⁷ Waldron explicitly says that the members of the idealised society he has in mind share “take rights seriously,” even though they disagree about what rights they have. “The Core of the Case,” p. 1365. The slogan is obviously borrowed from Ronald Dworkin, *Taking Rights Seriously*.

³⁸ The discussed the issue of reasonable disagreement in Chapter One.

³⁹ Waldron, *Law and Disagreement*, p. 245.

The Obligation of Subjects to the Sovereign is understood to last as long, and no longer, than the power lasteth by which he is able to protect them. For the right men have by Nature to protect themselves, when none else can protect them, can by no Covenant be relinquished.⁴⁰

Even on Hobbes's view, some outcomes are beyond the pale of acceptability. This insight, I believe, is what Waldron's account of democracy should be supplemented to recognise.

I have argued that political decision-making is no longer an "achievement" of the kind Waldron invokes when the result of decision-making is morally obnoxious. A democratic vote leading to slavery or oppression lacks the value that Waldron ascribes to it, for those whose moral status is flagrantly affronted by the decision cannot plausibly see the law as theirs—as the democratic polity's shared, provisional, reasonable policy out of the set of reasonable candidate policies.⁴¹ Substantive limits must be placed on the kinds of decisions we regard as tolerable.

§4. Toward a Foundational Judicial Review

I have argued that Dworkin's defence of judicial review, and Waldron's attack on it, are both too ambitious. Dworkin appears to affirm the view that *if* a case legitimately reaches a supreme or constitutional court, judges should approach the constitutionality of the statute or policy in question partly by drawing on their own philosophical convictions about the best interpretation of the moral language expressed in the relevant constitutional provisions. One crucial upshot of that approach is that some decisions about whether to permit or strike down a democratically enacted law will hinge on judges' determination that the law could not have been justified in accordance with their

⁴⁰ Hobbes, *Leviathan*, p. 272.

⁴¹ I make an abbreviated version of this point in "Punishment, Socially Deprived Offenders, and Democratic Community," *Criminal Law and Philosophy* 7, 1 (2013): 121-136.

preferred substantive conception of justice—even when it is plainly justifiable on the basis of some other reasonable conception of justice within the liberal family. Waldron is right to criticise this, and therefore to condemn such an expansive vision of judicial review, but his criticism is too sweeping for its purposes; he is not able to explain why judicial review would be inappropriate in cases of legislation *beyond the pale* of reasonable dispute.

We are thus in position to embrace the conclusion that a more restrictive judicial review oriented only to the preservation of the *foundations* of liberal justice (to the degree that those foundations are expressed in a constitutional text) could be justified. Call this

Foundational judicial review: A practice whereby judges are legally empowered to strike down legislation that could not be justified in accordance with any member of the family of reasonable liberal conceptions of justice.

Now the broad idea that justified judicial review should be restricted to fundamental cases is not, to be sure, original. The idea that justices should restrict themselves to clear and obvious instances of unconstitutionality remains a large part of judicial politics; the promise of exhibiting “deference” to legislatures is a familiar promise made to the U.S. Senate Judiciary Committee. In much of the early American experience with judicial review, judicial review was conceived not as an exercise in legal analysis, but as a political declaration from one branch of government to another that it had contravened a deep norm of the polity’s public philosophy.⁴² While Gerald Gaus and Thomas Christiano have both offered brief proposals along similar lines as part of their own theoretical frameworks,⁴³ the historical insight that judges should only condemn manifestly

⁴² Sylvia Snowiss, *Judicial Review and the Law of the Constitution* (New Haven: Yale University Press, 1990), p. 50. Another expression of this historical view is James B. Thayer, “The Origins and Scope of the American Doctrine of Constitutional Law,” *Harvard Law Review* 7, 3 (1893): 129–56.

⁴³ See Gaus, *Justificatory Liberalism* (New York: Oxford University Press, 1996), pp. 279ff; Thomas Christiano, *The Constitution of Equality* (New York: Oxford University Press, 2008), pp. 136ff. Christiano’s approach is more developed than Gaus’s; I explain later how I think it is misguided. Note that the idea that judicial review should operate only in a minority of cases is different from the idea that courts should ground their rulings on narrow legal grounds; for this view, see Cass

unreasonable legislation has not enjoyed a developed defence in contemporary Rawlsian theory. I have shown that this insight emerges as the plausible middle ground between the two prevailing views within the scholarly debate—those of Dworkin and Waldron. The rest of this section will provide my own preliminary defence of this insight within a Rawlsian framework, after which I will step back and consider its role within the wider theory of stability.

The central distinction between foundational judicial review and expansive judicial review is to be located in their respective accounts of *judicial virtue*. The expansive vision is best expressed by Rawls's discussion of a supreme court as "the exemplar of public reason." Though Rawls is agnostic about the necessity of judicial review in liberal democracies, he writes that if there *is* judicial review, judges who undertake it should "appeal to the political values they think belong to the most reasonable understanding of the public conception and its political values of justice and public reason."⁴⁴ This is precisely the ethic of judging that foundational judicial review condemns as illegitimate. Instead of asking whether a given law's constitutionality is vindicated by the judge's own preferred convictions about the relevant constitutional provisions, the judge should ask whether the law or policy could be deemed acceptable by *any* reasonable interpretation of those provisions, as indexed to any member of the family of reasonable (liberal) conceptions. On this modified version of judicial review, judges are granted far less "normative power" to alter citizens' and officials' reasons for action.⁴⁵ They still engage in interpretation of fundamental matters of principle, but the level at which they do so is far more circumscribed.

Sunstein, *One Case at a Time: Judicial Minimalism on the Supreme Court* (Cambridge: Harvard University Press, 2001).

⁴⁴ Rawls, *Political Liberalism*, p. 236.

⁴⁵ Joseph Raz, *Practical Reason and Norms* (Oxford: Oxford University Press, 1999), p. 98

To elaborate this ethic of judging, let us consider three examples. Firstly, consider the example of hate speech. Suppose a plaintiff comes to court with the dispute that a local law banning hate speech has violated her fundamental rights. On my view, assuming (1) that hate speech *is* a subject about which there is reasonable disagreement, (2) that there is a constitutionally entrenched right to free speech, and (3) that the law in question was indeed democratically enacted, judges should defer to the democratic decision and permit the law to stand. Suppose, however, that a plaintiff comes to court with the claim that a law has criminalised her communication of her beliefs about electoral reform. On this point, the matter is clear: the law should be struck down as unconstitutional, because there cannot be reasonable disagreement about its injustice.⁴⁶

Secondly, consider the example of religious liberty. A law that criminalises adherence to a particular religion clearly imperils citizens' standing as free and equal, and should therefore be struck down by courts. So, too, should a law that permits freedom of conscience in private, but nevertheless requires public affirmation of a state religion in political affairs, be invalidated; there is no plausible argument from the political liberal conception of moral personality of how this could be acceptable. Nevertheless, it remains an open question whether citizens should receive exemptions from generally applicable laws when those laws burden their religious practices. The degree of schizophrenia on this question by the U.S. Supreme Court over the last century is matched only by the degree of profoundly reasonable disagreement among political

⁴⁶ For general discussion of free speech, see T. M. Scanlon, "A Theory of Freedom of Expression," in *The Difficulty of Tolerance: Essays in Political Philosophy* (Cambridge: Cambridge University Press, 2003). For the specific contention that hate speech legislation is condemned by the most reasonable conception of justice, and thus could not be covered by the most reasonable understanding of freedom of speech (albeit not argued for in those Rawlsian terms), see Jeremy Waldron, *The Harm of Hate Speech* (Harvard: Harvard University Press, 2012). For an argument in opposition, see Dworkin, *Freedom's Law*, pp. 223-26.

philosophers on the matter; this should be instructive.⁴⁷ In such a case, justices should not impose their own convictions about justice on all, but rather let the body of deliberating citizens decide.

Finally, consider the example of freedom of contract. American critics of judicial review who avow a liberal-egalitarian political theory recurrently invoke the numerous cases associated with *Lochner v. New York* (1905) as grounds for suspicion about judicial review.⁴⁸ Now I believe there can be reasonable disagreement in liberal political philosophy about the scope of freedom of contract.⁴⁹ But on no reasonable interpretation of the provision of freedom of contract—as a moral right that hangs together with the rest the rights we owe each other as free and equal—is it appropriate to conclude that the democratic community, especially in circumstances of distributive injustice, is not entitled to take reasonable measures that reduce the potential for exploitation of the poor.⁵⁰

⁴⁷ For the pro-exemption position, see the case *Sherbert v. Verner*, 374 U.S. 398 (1963) and the argument in Martha Nussbaum, *Liberty of Conscience: In Defense of America's Tradition of Religious Equality* (New York: Basic Books, 2008). For the anti-exemption position, see the case *Employment Division v. Smith*, 494 U.S. 872 (1990) and the argument in Brian Barry, *Culture and Equality* (Cambridge: Harvard University Press, 2001), 155-93.

⁴⁸ This case struck down a law limiting the number of hours bakers could work to ten on the grounds that it violated their freedom of contract. *Lochner v. New York*, 198 U.S. 45 (1905).

⁴⁹ For a recent defence of the claim that Rawlsians have neglected the issue of freedom of contract, see John Tomasi, *Free Market Fairness* (Princeton: Princeton University Press, 2012); for criticism, see Samuel Arnold, "Right-Wing Rawlsianism: A Critique," *Journal of Political Philosophy* (forthcoming): available online at <http://onlinelibrary.wiley.com/doi/10.1111/jopp.12001/abstract>, accessed 15 April 2013.

⁵⁰ A quite different minimalist of judicial review is undertaken by Thomas Christiano, according to which judicial review should only protect "the core" of liberal rights. I nevertheless believe that Christiano misidentifies how to identify what sorts of cases are protected as part of "the core" of liberal rights, making his proposal inadequate. His strategy is to triangulate between, firstly, the right-grounding interests, secondly, the right itself, and thirdly, an action that is disallowed by the law in question, and identify the directness or strength of the connection between the right-grounding interests and the action. If it is a strong connection, the action is protected as part of the right's core, and if there is not, the action falls to the judicially unprotected "periphery." But while some reasonable disagreements about rights may have this character, I believe it is a mistake to use this strategy as the approach for determining whether a given disagreement about justice is reasonable. For, it would fail to mark out some of the most pressing instances of reasonable disagreement within political liberalism as reasonable

While I have drawn my comments from American cases, foundational judicial review is an ideal for judicial review it would operate in a well-ordered liberal democracy. As such, the proposal *itself* is not anchored in any *existing* constitutional text. Thus I have avoided the problem of what judges should do when certain constitutional specifications are themselves morally undesirable elements of a constitution. But I do not thereby wish to suggest that foundational judicial review should be *legally unmediated* in the manner that political philosophers often talk of judicial review as being—as though judges were moral gatekeepers who weigh in on public policy *whenever* they perceive that a fresh piece of legislation has gone off the moral cliff. While there would be virtues to such an approach—enabling judges to reason about the first-order questions directly without having to engage in highly distorted and often implausible legal reasoning⁵¹—it would be misconceived. The distinction between the reasonable and the unreasonable is extremely helpful for philosophers interested in framing a general outlook. But the actual process of vindicating a particular policy as reasonable is another matter. Political philosophers

disagreements. Consider the case of whether religious freedom demands that minorities burdened by generally applicable laws receive exemptions from religious practices. For those, like Martha Nussbaum, who believe that justice *demand*s such exemptions, the connection between the right-grounding interests, on the one hand, and the actions, on the other, is extremely strong. And yet—if I am right—this issue is subject to reasonable disagreement. Brian Barry, for example, believes that it is fundamentally *incompatible* with a vision of liberal order to conceive of the law as issuing discrete demands to discrete citizens on the basis of their group identity. But the disagreements to be had between defenders of Nussbaum’s position, on the one hand, and defenders of Barry’s, on the other, are not properly deemed reasonable, and thus open to democratic determination rather than judicial determination, *because* they concern putatively peripheral cases. Nussbaum would argue that the exemptions are central to the expression of her protagonists’ religious identity. Rather, the determination that the disagreement is reasonable is made because of the difficulty of the subject matter, the compellingness of the arguments on both side, and the seemingly central role of *judgement* in adjudicating who is right. At any rate, Christiano’s proposal is indexed to his distinctive view of liberal political philosophy, which departs from political liberalism in numerous respects that I cannot discuss here. See Christiano, *The Constitution of Equality*, pp. 136ff.

⁵¹ For example, it is extremely morally unhealthy for a democracy to confront the fundamental question of whether we morally owe each other basic healthcare by addressing whether the Article I power to regulate interstate commerce extends to regulation of the healthcare market. *National Federation of Independent Business v. Sebelius*, 567 U.S. (2012).

should not be surprised that existing practices of judicial review are indexed to particular legally enumerated rights. There is a strong moral presumption in favour of indexing judicial reasoning to some legislatively enacted bill of rights, be it constitutionally entrenched in some manner or not.⁵²

Two further comments are necessary. Firstly, the defence of foundational judicial review does not presuppose judicial infallibility. Why are we to think that if ordinary citizens are so fallible, that judges will be any better? The historical record on judicial review in the U.S. is mixed on this point, making an historical evaluation of judicial review's moral success hopelessly indeterminate. That the historical record is mixed, I believe, is precisely because of the prevailing ethic that I seek to condemn. Because the account of judicial review I have presented is in keeping with certain skills of which the judicial profession has proven itself capable,⁵³ and is vindicated as possible by the fact that there *have* been numerous instances of judicial review marshalled for the benefit of justice, it is enough for the proposal to have plausibility. Its real test is whether the law schools and legal profession will be able to cultivate the culture of restraint and epistemic modesty that I have argued is essential to their job. But that test cannot be completed by normative political theorists.

Secondly, foundational judicial review does not aspire to *eliminate* disagreement among judges about what courses of action are best. While a well-ordered liberal democracy will feature sufficient unity of opinion about the sorts of laws that belong to

⁵² There is, of course, tremendous controversy over what it means for something to be constitutionally entrenched. See, for example, John Gardner, "Can There Be a Written Constitution?" in *Oxford Studies in the Philosophy of Law*, Volume 1, edited by Leslie Green and Brian Leiter (New York: Oxford University Press, 2011).

⁵³ For defence of the idea that constitutional court judges are engaged in the practice of applied political theory, see David Robertson, *The Judge as Political Theorist: Contemporary Constitutional Review* (Princeton: Princeton University Press, 2010). For scepticism that judges are best positioned to engage in moral reasoning, see Jeremy Waldron, "Judges as Moral Reasoners," *International Journal of Constitutional Law* 7, 1 (2009): 2-24.

its “black list”—the sorts of laws that are beyond the pale of reasonable disagreement—there are bound to be hard cases (e.g., wiretapping in times of grave national emergency), although a case’s status as difficult is itself a reason to think that the considered decision of the citizenry ought to be honoured. But while foundational judicial review does not succeed in eliminating disagreements about whether judicial opinions were justified, it aspires to *relocate* them. The relocation for which I have argued in this section is a *substantive* relocation—concerning the “what” of judicial review. In the next two sections, I shall complete my elaboration of foundational judicial review by arguing for a (partially) *temporal* relocation, as well, addressing the philosophically neglected question of *when* in the political process judicial review ought to transpire, and why it matters.

§5. Judicial Review and the Impairment of Civic Relations

“Legislation deserves respect,” writes Waldron, “because of the achievement it represents in the circumstances of politics: action-in-concert in the face of disagreement.”⁵⁴ I have argued that our sense of democratic rule as an “achievement” becomes perverse if democracy is marshalled to arrive at some deeply unjust outcome, and thus that democracy’s outcomes are only “achievements” if they fall within some morally tolerable set. According to an important school of thought in democratic theory, however, this is too fast. The mere fact that a democratic decision generates an unjust outcome is not sufficient for us justifiably to conclude that judicial review is appropriate. “To determine when judicial review is justifiable,” Corey Brettschneider writes, “we must balance two duties: first, the duty to uphold the substantive values [freedom and equality] which underlie democratic procedures and, second, the duty that comes from the fact

⁵⁴ Waldron, *Law and Disagreement*, p. 108.

that a law was passed by a democratic procedure.”⁵⁵ Such balancing is necessary because “*the democratic authority of a law still retains some normative force even when it undermines these substantive values [of liberal justice]*.”⁵⁶ We must “balance” the disvalue of rights-violations with the intrinsic disvalue of disrespecting citizens’ status as co-rulers.

This idea of balancing the costs of judicial review is, I believe, seriously misconceived, but an analysis of why it is misconceived yields an important insight. Consider three versions of the thesis that “democracy is...a *pro tanto* good,” and so while it is “thereby defeasible in the course of practical deliberation,” overriding it is always grounds for regret.⁵⁷ Firstly, democracy instantiates what Thomas Christiano calls an ideal of “public equality”—it honours citizens’ status as equals in the face of their intractable disagreements.⁵⁸ Secondly, it realises a form of procedural fairness, as it is not designed in a way rigged in favour of any particular person’s view.⁵⁹ And thirdly, it intrinsically honours our liberal self-understanding of people capable of choice; this is the argument for democracy’s intrinsic value endorsed by Waldron. If one or all of these goods are *pro tanto* valuable,⁶⁰ then it is plausible to think that judicial review *does* generate costs, by disrespecting the *pro tanto* value of democracy. Depending on the degree of the injustice, the *pro tanto* value of democracy may even be weighty enough to *override* the democratic decision.

But notice a disturbing corollary of this position, which I mentioned briefly in Chapter One. If democracy is intrinsically valuable, then a democratically perpetrated

⁵⁵ Brettschneider, “Judicial Review and Democratic Authority: Absolute vs. Balancing Conceptions,” *Journal of Ethics and Social Philosophy* (2011): 1-9, p. 3.

⁵⁶ *Ibid.*

⁵⁷ See Shelly Kagan, *The Limits of Morality* (Oxford: Clarendon Press, 1989), p. 17.

⁵⁸ Christiano, *The Constitution of Equality*, pp. 70ff.

⁵⁹ For an argument as to why procedural fairness is always a *pro tanto* good, see John Broome, “Fairness,” *Proceedings of the Aristotelian Society* 91, 1 (1991): 87-101.

⁶⁰ On the nature of a *pro tanto* wrongful action, see Shelly Kagan, *The Limits of Morality* (Oxford: Clarendon Press, 1989), p. 17.

injustice is somewhat less unjust than an injustice ordered by an authoritarian.⁶¹ That, as Eric Beerbohm rightly says, does not match our lived experience. “South Africa’s partially democratic machinery behind apartheid didn’t make the system mildly morally better.”⁶² It made it worse. That an egregious injustice was not the one-off whim of some insane madman holding the reins of state power, but rather an organised, systematic undertaking “made through an impeccably democratic process,” exacerbates our anger.⁶³ Democracy *simpliciter* cannot be *pro tanto* valuable, since if it were, such value would “take the edge” off democratic injustice.

This, however, is implausible. Consider a collective decision about which minority group to annihilate. Suppose that everyone gets to vote, even in the members of the minorities. It would be outrageous to suppose that there is even a scintilla of moral value in making the decision in such a manner. H. L. A. Hart notes that “the germ at least of justice” is present in “the most odious laws” when they follow the strictures of the rule of law. But this cannot be right.⁶⁴ (Indeed, an unpredictable lapse from otherwise consistent genocide marks the *advance* of justice, not the retreat of it.)

Brettschneider’s defence of balancing cannot be sustained.⁶⁵ It does not therefore follow, however, that judicial review does not involve moral loss in some sense—as

⁶¹ Beerbohm, *In Our Name*, p. 25.

⁶² *Ibid.*, p. 36.

⁶³ *Ibid.*

⁶⁴ See the discussion in Leslie Green, “The Germ of Justice,” *Oxford Legal Studies Research Papers No. 60/2010*, available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1703008, accessed 20 March 2013. Cf. See Lon Fuller, *The Morality of Law* (New Haven, Yale University Press, 1964), Chapter Four.

⁶⁵ Even if it could, Brettschneider does not explain how this balancing should transpire. He says: “while there always is a loss to the democratic ideal when democratic procedures are overturned, at times the loss to democracy is greater if these procedures are not overturned” ((*Democratic Rights*, pp. 138ff). He suggests that the U.S. Supreme Court ought to strike down all laws that violate a basic right, but that on issues that are not rights-implication, legislatures ought to decide. But there will be reasonable disagreement about *what counts* as “rights-implicating,” and justices will be faced with a crucial question of democratic theory—which Brettschneider does not consider (with the exception of welfare rights, which he regards as best left to legislatures for

though we should be morally indifferent between the determination of some vital policy matter by the deliberating citizenry and its determination by judges on a constitutional court (an indifference that Dworkin occasionally suggests). The loss I have in mind is the intrinsic disvalue of the democratic choice to enact egregious injustice, a choice that generates an impairment in citizens' relationships with one another. But *judicial review* does not generate that loss: people's wrongful choices do. Judicial review is *occasioned* by lamentable circumstances, but it does not *cause* those lamentable circumstances. Brettschneider and defenders of similar approaches mistake the medicine for the poison.

The idea that democracy has noninstrumental value, as I argued in Chapter One, is best understood as the claim that democratic procedures empower people to respect one another's status as free and equal. It is those *choices* that have enduring intrinsic value, as expressions of citizens' affirmations of their relations with others as free and equal—not some fixed feature of the democratic machinery itself, like the fact that it only accepts one ballot per person. But the decision to enact egregious injustice is not a decision of that intrinsically valuable kind at all.

§6. Judicial Review's Temporal Dimension

On the story I have told, judicial review does not trigger moral loss; rather, moral loss triggers it. In this way, it seems, judicial review is fated always to come too late. Citizens' relationships will have already been impaired. Judicial review blocks the (continued) manifestation of morally pernicious or ignorant attitudes in wrongful policies; but much of the damage has already been done merely by the way members of the targeted

empirical reasons)—of how judges should decide on issues whose justice is subject to reasonable disagreement. Moreover, insofar as it is conceivable that *all* the demands of justice could be cashed out in terms of individual rights, how are we to decide which rights are sufficiently important to trigger judicial review?

minority, for example, must now—lest they suffer delusion—understand the character of their relations with their fellow citizens. If circumstances arise that render judicial review appropriate, members of that minority—albeit provisionally protected by courts—can only conclude: we are surrounded by enemies.

This state of affairs is not the mark of a well-ordered liberal democracy. It thus appears that my account of judicial review is only suited to help correct for serious *deviations* from our ideal social order. Like a SWAT team—well-trained and waiting in the wings for the day that some horrific mass shooting transpires—the judges of the constitutional court sit quietly,⁶⁶ fingers-crossed that their co-citizens will never need to avail the judges of their services. Without denying that judicial review ought to serve this function, I want to ask: can we hope for something more constructive, more promising, from judicial review? The very need for judicial review, as I have described it, could of course be obviated simply if people would accurately identify and choose to comply with their fundamental duties of justice in the first place. But might judicial review—understood in a more capacious, institutionally imaginative form—help us realise that very aspiration? In the remainder of this chapter, I shall sketch and defend an affirmative answer to those questions.

It is important to see that judicial review, as currently practiced, is not merely an institution that silently blocks democratic action. Rather, it enables judges to articulate their reasons for deciding as they have—explaining how the fundamental constitutional values, properly interpreted, yield the determination that the court has reached. Ronald

⁶⁶ At least in their capacity as participants in rights-based judicial review of legislation, anyway. Due to spatial constraints, I simply cannot address the matter of whether rights-based judicial review should be part of the normal proceedings of the judiciary, or whether a special constitutional court is institutionally best. I aspire to connect these debates to issues in constitutional theory in later work, but spatial constraints restrict me to the literature of normative political theory.

Dworkin continually emphasises not merely the narrow benefits of judicial review in correcting this or that majoritarian mistake, but in awakening and nourishing a political culture in which large questions of value are taken seriously. By providing opportunities for citizens to listen to Supreme Court hearings and await the news of a major new court decision, judicial review, according to Dworkin, didactically infuses the democratic political culture with a passion for justice:

When an issue is seen as constitutional...and as one that will ultimately be resolved by courts applying general constitutional principles, the quality of public argument is often improved, because the argument concentrates from the start on questions of political morality...When a constitutional issue has been decided by the Supreme Court, and is important enough so that it can be expected to be elaborated, expanded, contracted, or even reversed by future decisions, a sustained national debate begins, in newspapers and other media, in law schools and classrooms, in public meetings and around dinner tables. That debate better matches [a] conception of republican government, in its emphasis on matters of principle, than almost anything the legislative process on its own is likely to produce.⁶⁷

Waldron finds this laudation laughable:

A star-struck people may speculate about what the Supreme Court will do next on abortion or some similar issue; they may even amuse each other, as we law professors do, with stories of how we would decide, in the unlikely event that we were elevated to that eminent tribunal. The exercise of power by a few black-robed celebrities can certainly be expected to *fascinate* an articulate population. But that is hardly the essence of active citizenship.⁶⁸

How, Waldron is asking, could judicial review *morally empower* citizens in the way Dworkin intimates if its *modus operandi* is that of *bypassing* citizens' moral choices? This exchange, however, fuels an unhealthily *binary* attitude in our institutional imagination toward judicial review. There is a third way: a scheme in which judicial review both teaches *and* empowers.

⁶⁷ Dworkin, *Freedom's Law*, p. 345.

⁶⁸ Waldron, *Law and Disagreement*, p. 291. Emphasis is Waldron's.

Excessive focus on legislation reviewed by U.S. Supreme Court has resulted in an institutionally unimaginative debate on judicial review in normative political theory.⁶⁹ The chain of events is familiar: majoritarian legislation is challenged by a putatively burdened citizen who takes the government to court to decide whether to strike down the legislation or not.⁷⁰ But judicial review does not everywhere have this *ex post* character. In some jurisdictions, the place of judicial review is *ex ante*: a panel of judges reviews the proposed legislation's content prior to the time of legislative decision. There is considerable institutional variability even within the *ex ante* model—between systems in which judges are legally empowered to block the occurrence of a vote, and systems in which their services are merely advisory; and between systems which a court itself determines whether to undertake review, or whether an official in another branch can demand such review.⁷¹ Waldron doubts whether these cases really count as judicial review at all, but I shall resist such debate-ending terminological fiat.⁷² We are still talking about a judicial body, comprising experts trained in the law and well-educated in the constitutional values, reviewing legislation; the only difference is whether it reviews it *before* or *after* the vote.

⁶⁹ Lawyers and constitutional theorists are not nearly as guilty of this charge. See Mark Tushnet, *Taking the Constitution Away from the Courts* (Princeton: Princeton University Press); Sanford Levinson, *Constitutional Faith* (Princeton: Princeton University Press, 1988).

⁷⁰ This is a simplification, but it serves our purposes.

⁷¹ A system along these lines has been in operation in many U.S. states, whose state courts have powers to offer prospective “advisory opinions” about impending legislation or executive actions. The general idea that there be a dialogue between legislatures and judiciaries has attracted significant attention in constitutional theory and law. See, for example, Kent Roach, “Common Law Bills of Rights as Dialogue Between Courts and Legislatures,” *The University of Toronto Law Journal* 55, 3 (2005): 733-66. I regrettably cannot delve into the more technical dimensions of constitutional theory and practice here.

⁷² Waldron, “The Core of the Case,” p. 1359. In the United Kingdom, for example, the UK Supreme Court can issue a non-binding “declaration of incompatibility” when it notices an inconsistency between some Act of Parliament and the Human Rights Act. If a minister so chooses, (s)he can even decide to trigger a re-vote on the legislation.

Let us imagine awakening to find ourselves in a well-ordered liberal democracy. While we reasonably disagree about justice, fundamental justice is secured. We realise, however, that we are fallible—in precisely the ways I described in Chapters Two and Three—and we are concerned to take proactive measures to mitigate the possibility of our fallibility. However, we aspire to sustain the character of our just social relationships, and so we are reluctant to endorse mechanisms that secure just outcomes by bypassing our first moral power—at least not as the first course of action. We aspire, first and foremost, to *fortify* our first moral power; we want to endorse mechanisms that *assist us* in choosing to do what is right, as opposed to take the decision out of our hands. My proposition is that we would create a specially designed constitutional court whose primary function would be to conduct “*ex ante* weak foundational judicial review,” according to which it would evaluate pending legislation for its compatibility with any reasonable interpretation of the constitutional values and, in the eventuality that an incompatibility was determined, present its arguments for this incompatibility to the legislature prior to the vote, publicly fielding questions from legislators and citizens about their reasoning. (A variant on this view would empower the court to take the legislation out of the legislature and make it a referendum.⁷³) The legislature (or citizens) would subsequently hold a vote on the legislation. We could imagine a similar process by which judges—or citizens—could flag complaints of unconstitutionality about longstanding pieces of legislation; rather than strike old legislation immediately down upon believing it to be beyond the pale, judges would issue a ruling, triggering a similar deliberative session

⁷³ I shall not defend that innovation here. To identify and defend its distinctive appeal rigorously would necessitate endorsing a certain theory of the citizen-legislator relationship, an extremely important and, I believe, neglected topic in contemporary normative democratic theory, but one that lamentably transcends the scope of this dissertation. For an analysis of this relationship, see Beerbohm, *In Our Name*, pp. 193-225. See also the classic discussion in Edmund Burke, “Speech to the Electors of Bristol,” *The Works of the Right Honourable Edmund Burke, Vol. 1* (London: Henry G. Bohn, 1854), pp. 446ff.

and a vote. Thus judicial review, rather than bypassing citizens' sense of justice, could serve as a general fortifying practice oriented to *enabling* citizens and their officials to recognise the pending legislation as unjust. Such a vision of *ex ante* weak foundational judicial review is fully compatible with citizens' normative self-understanding and serves to fortify democratic officials and citizens in their quest to make just choices.

Consider the following problem, however. What would happen if officials or citizens nevertheless chose to vote for legislation that was certified as "beyond the pale" during *ex ante* review? In light of what I have argued in this chapter, we have every reason to think that *ex post* strong foundational judicial review would thereby become lamentably appropriate. Indeed, I shall now argue that citizens of a well-ordered liberal democracy would have reason to install such a system in advance, for the purposes of correcting for egregious injustice in moments when they are led morally astray. By conjoining *ex ante* weak foundational judicial review with a back-up mechanism—*ex post* strong foundational judicial review—such citizens properly *precommit themselves* to be faithful to their own values in moments when democracy morally malfunctions.

The idea of precommitment has been invoked in debates on judicial review before.⁷⁴ John Potter Stockton sums up the traditional understanding of the matter: "Constitutions are chains with which men bind themselves in their sane moments that they may not die by a suicidal hand in the day of their frenzy."⁷⁵ These arguments have been subjected to withering criticism. Jon Elster—a onetime defender of the precommitment defence of judicial review—retracts his earlier position that elections

⁷⁴ Jon Elster, *Ulysses and the Sirens* (Cambridge: Cambridge University Press, 1979); Stephen Holmes, *Passions and Constraint* (Chicago: University of Chicago Press, 1995); Sunstein, "Constitutionalism and Secession," *University of Chicago Law Review* 58 (1991): 633-70, p. 641; Samuel Freeman, "Constitutional Democracy and the Legitimacy of Judicial Review," pp. 353-54.

⁷⁵ Cited in J.E. Finn, *Constitutions in Crisis* (New York: Oxford University Press, 1991), p.5, which is further cited in Elster, *Ulysses Unbound* (Cambridge: Cambridge University Press, 2000), p. 89.

“can be interpreted...as the electorate’s method of binding itself and of protecting itself against its own impulsiveness,”⁷⁶ insisting:

Obviously, no electorate did anything of the kind. If voters lack the means to recall their representatives at will, the explanation surely has more to do with the motives of politicians than with those of voters.⁷⁷

Elster sums up the discovery that changed his mind: “In politics, people never try to bind themselves, only to bind others.”⁷⁸ But defenders of the precommitment view, such as Samuel Freeman and Stephen Holmes, invite these criticisms by casting the precommitment view as an *explanation* of what judicial review *is*, as though some actual moment of “self-binding” transpired. Judicial review in the United States was the historical outcome of a power struggle between particular politicians. But there is a distinctive use of the idea of precommitment that is more plausible, and indeed applies with great force to the vision of judicial review I am defending: *hypothetical* precommitment. What matters for my view is *not* that we actually *did* choose to bind ourselves to strong judicial review. What matters is that we have *duties* to precommit. Accordingly, if we find ourselves subject to an on-going scheme of judicial review, then—so long as it is operating in the way it ought to—we have a duty not to dismantle it, but to sustain it. No system of judicial review operates without the continued support of citizens and officials who dictate the terms of its practice and adhere to its results. Likewise, if we find ourselves in a circumstance in which we would have occasion to *create* judicial review, we would have duties to do so. That is the grain of truth in the precommitment proposal: it is an arresting way of expressing the fact that we have *moral reason* to embrace judicial review.

⁷⁶ Elster, *Ulysses and the Sirens*, p. 91, retracted in *Ulysses Unbound*, p. 90.

⁷⁷ Elster, *Ulysses Unbound*, p. 90.

⁷⁸ *Ibid.*, p. ix. Elster attributes the quotation to Jens Arup Seip, secured from personal conversation.

I thus conclude that citizens ought to endorse a hybrid proposal, combining *ex ante* weak foundational judicial review—as a fortificational measure of which democratic citizens and officials avail themselves in order to keep their democratic reasoning and decision-making on track—and *ex post* strong foundational judicial review—to be used rarely as a prophylactic that protects vulnerable minorities from moral catastrophe. Let me close by making some orienting comments on this hybrid proposal. Firstly, while *ex ante* judicial review clearly belongs to “ideal theory” as I described it at the end of Chapter Two—it belongs to our ideal of well-ordered human relationships with one another as one of the many fortificational measures upon which we draw to fortify our first moral power—the place of *ex post* judicial review within the theory is more precarious; for, the circumstances that make *ex post* judicial review appropriate indicate that we have deviated from our ideal just social relationships, and thus find ourselves in circumstances to which non-ideal theory, by definition, applies. I propose that we conceptualise *ex post* judicial review as part of an “upper tier” of non-ideal theory, comprising a package of mechanisms (not all of which are fortificational in character) that sit unused in a well-ordered society but become appropriate in moments of *deviation* from that ideal. These are mechanisms we design to help steer our institutions back toward justice when they have fallen away—like the SWAT team I described earlier, waiting in the wings and prepared to enforce justice. Such measures may not be fortificational in character; the SWAT team’s method does not aspire to increase the free success of prospective wrongdoers’ first moral power, but to bypass it through threats of sheer force. Likewise, the *modus operandi* of *ex post* strong judicial review is to bypass citizens’ democratic agency. Nevertheless, we can envision (indeed, Dworkin *does* envision) an *ex post* strong judicial review that *does* have fortificational effects, and it is plausible to think that judges who find themselves in such circumstances ought to

undertake *ex post* judicial review in the manner that maximises those effects. We can conclude that *ex post* judicial review thus *partly* belongs to the theory of stability⁷⁹ insofar as elements of its content and structure are justified with reference to fortificational concerns, whereas *ex ante* weak judicial review serves as a paradigmatic element of an ideal society's fortificational infrastructure.

The second concern I seek to address concerns the relation *between* the two forms of judicial review. Some are bound to argue that this *ex post* form of judicial review—albeit conceived as merely a safety mechanism—will render pointless the *ex ante* form of judicial review I specified and defended earlier in this section. Indeed, if the legislature knows that the pronouncement of the *ex ante* court will be swiftly implemented to overrule the democratic edict if the legislature does not alter its stance, it may be difficult for legislators to construe the *ex ante* pronouncement as anything other than a *threat*: “Do not dare vote for this, or we will strike it down.” This is legitimate worry, but we can accommodate it with two proposals and a reminder. The first proposal is that *ex post* strong judicial review need not be immediate. The court could wait until the legislation was enacted, and only after the initiation of a complaint by a burdened citizen would it then take up the issue again. This proposal has the added benefit of providing an opportunity for the judges themselves to reflect on whether they may have been mistaken in deeming the legislation “beyond the pale” of reasonable disagreement. The second proposal is to pursue a judicial division of labour, having separate bodies of

⁷⁹ Some may dispute my idiosyncratic definition of what “the theory of stability” involves. Why not define it more capaciously, to refer to *all* the mechanisms that we properly design and subject to ourselves—even if they involve agency-bypassing—to make sure justice’s “bottom line” (i.e., protecting others’ entitlements) is achieved? The answer is that this would forfeit what is distinctive about stability on both Rawls’s view and my re-tooled variation of it: the *freely willed* enduring maintenance of justice. The theory of stability, as I conceive it, applies to this distinctive concern: how we achieve justice, and sustain it, simply through the exercise of our practical reason in the world.

officials undertake each element of the process. So, consider the contemporary United States. The most straightforward path to implement my hybrid proposal there would be for the existing courts to adopt my minimalist ethic of judicial review in undertaking their *ex post* strong exercise of it—striking down enacted legislation only when it is far beyond the pale—but creating a new institution to offer *ex ante* weak review of pending legislation to search for incontrovertible incompatibilities with the constitutional values. Taken together, these two proposals short-circuit the worry that legislators are bound to construe weak review as a threat. That they *not* construe as a threat is very important, and precisely because the point of weak judicial review is to provide legislators with an opportunity to stop, think, and wonder whether they really are doing what is right, or whether they are acting opportunistically, maliciously, or ignorantly. As fallible, we all make mistakes, and if we are paused before we make them and enjoined to reconsider—if we are given the opportunity to get it right before we pull the trigger—we need not regard the integrity of our social relations as compromised. Rather, through working together, it will have simply been fortified in the face of our fallibility.

CHAPTER FIVE

The Fortificationist Theory of Punishment

In this chapter I aim to describe the proper role of criminal punishment in a well-ordered liberal democracy. My motivating conviction is that punishment in an ideal society is not the infliction of suffering upon those who deserve it,¹ and it is not the removal of a benefit unfairly gained;² nor is it the retaliation justified by a natural right to self-defence,³ nor the hard treatment that is putatively necessary for the state to “express” its disapproval.⁴ The practice of punishment, instead, belongs to the theory of stability: it is a practice that citizens have a fortificational duty to design and maintain—and, in the event of a criminal choice, submit to—to fortify their first moral power in the aftermath of its defective operation.

Political liberals have tended to neglect the theory of punishment.⁵

Notwithstanding the Rawlsian mantra that political power is the “coercive power of free

¹ Michael S. Moore, *Placing Blame: A General Theory of the Criminal Law* (New York: Oxford University Press, 1997). See also Lawrence H. David, “They Deserve to Suffer,” 32, 4 (1972): 136-40.

² Richard Dagger, “Playing Fair with Punishment,” *Ethics* 103 (3) (1993): pp. 473-488.

³ Daniel Farrell, “The Justification of Deterrent Violence,” *Ethics* 100, 2 (1990): 301-17, “On Threats and Punishments,” *Social Theory and Practice* 15, 2 (1989): 125-54, “Punishment Without the State,” *Noûs* 22, 3 (1988): 437-53, “The Justification of General Deterrence,” *The Philosophical Review* 94, 3 (1985): 367-94; Warren Quinn, “The Right to Threaten and the Right to Punish,” *Philosophy and Public Affairs* 14, 4 (1985): 327-73; and in response to Quinn, Michael Otsuka, “Quinn on Punishment and Using Persons as Means,” *Law and Philosophy* 15, 2 (1996): 201-08.

⁴ Joel Feinberg, “The Expressive Function of Punishment,” in *Doing and Deserving* (Princeton: Princeton University Press, 1970).

⁵ Two exceptions are Corey Brettschneider, *Democratic Rights*, pp. 96-113, and Emmanuel Melissaris, “Toward a Political Theory of Criminal Law: A Critical Rawlsian Account,” *New Criminal Law Review* 15, 1 (2012): pp. 122-55. Critical discussions of Rawls and punishment have occurred in Bonnie Honig, “Rawls on Politics and Punishment,” *Political Research Quarterly* 46, 1 (1993): 99-125; Matt Matravers, “Mad, Bad or Faulty? Desert in Distributive and Retributive Justice,” in *Responsibility and Distributive Justice*, edited by Carl Knight and Zofia Stemplowska (New York: Oxford University Press, 2011); and Samuel Scheffler, “Justice and Desert in Liberal Theory,” in *Boundaries and Allegiances: Problems of Justice and Responsibility in Liberal Thought* (New York: Oxford University Press, 2001). For the argument that political liberalism is doomed not to be able to justify punishment, see Stanley C. Brubaker, “Can Liberals Punish?,” *The American Political Science Review* 82, 3 (1988): 821-36. For the argument that the problem of punishment inspires productive revisions in liberal theorising, see Matt Matravers, *Justice and Punishment: The*

and equal citizens,”⁶ it is surprising just how little work has been done to see how a theory of punishment might fit within the most prominent normative political theory of the twentieth century. This is partly the result, I suspect, of Rawls’s influential contrast between ideal theory and non-ideal theory, and the fact that one of the simplifying assumptions Rawls employs when constructing the former is an assumption of “strict compliance” with whatever principles are selected. Such an idealising⁷ assumption allows parties to the original position “to imagine the results of getting ‘up and running’ the institutions embodying different conceptions of justice...”⁸—thereby enabling easy comparison. But Rawls’s rightful insistence that ideal theory is a philosophically prior task to non-ideal theory—since it is “the only basis for the systematic grasp”⁹ of non-ideal problems¹⁰—may have had the consequence of motivating fewer political liberals to turn to the vexed questions of partial compliance. As a result, political liberalism lacks a distinctive theory of punishment. This chapter offers an introductory sketch of an account intended to fill that lacuna.

Rationale of Coercion (New York: Oxford University Press, 2000), and Thaddeus Metz, “How to Reconcile Liberal Politics with Retributive Punishment,” *Oxford Journal of Legal Studies* 27, 4 (2007): 683-705.

⁶ Rawls, *Political Liberalism*, pp. 139.

⁷ See Onora O’Neill, “Abstraction, Idealization and Ideology in Ethics,” in *Moral Philosophy and Contemporary Problems*, edited by D.G. Evans (Cambridge: Cambridge University Press, 1988), pp. 55-69; *Towards Justice and Virtue* (Cambridge: Cambridge University Press, 1996), esp. pp. 38-44. For incisive discussion on O’Neill’s distinction between idealisation—the deliberate assumption of falsehood—and abstraction—the deliberative simplification of a phenomenon for the purposes of analysing it—see Zofia Stemplowska, “What’s Ideal About Ideal Theory?” *Social Theory and Practice* 34, 3 (2008): 319-40.

⁸ A. J. Simmons, “Ideal and Nonideal Theory,” *Philosophy & Public Affairs* 38, 1 (2010): 5-36, p. 9. See also Michael Phillips, “Reflections on the Transition from Ideal to Non-Ideal Theory,” *Notus* 19, 4 (1985): 551-70, pp. 553-56.

⁹ Rawls, *A Theory of Justice*, p. 8.

¹⁰ For criticism of this view, see Amartya Sen, *The Idea of Justice* (Cambridge: Harvard University Press, 2009). For related but more radical criticisms, see Colin Farrelly, “Justice in Ideal Theory: A Refutation,” *Political Studies* 55 (2007): 844-64; Charles W. Mills, “‘Ideal Theory’ as Ideology,” *Hypatia* 20, 3 (2005): 165-84.

In Section 1, I briefly survey philosophers' persisting quest to specify a conception of punishment that aims to prevent crime in manner consistent with respect for criminal offenders as moral agents. In Section 2, I sketch the phenomenology of a reasonable citizen in the aftermath of criminal failure, arguing that the morally appropriate attitude of such an agent—in addition to apologising to and compensating her victims—should be to identify why she permitted herself to break the particular legal norm she broke, and to resolve to reduce the likelihood that she will let herself contravene that norm again. On the basis of this phenomenology, in Section 3 I defend the proposition that criminal offenders have special fortificational duties to undertake potentially demanding extra measures to shore up their first moral power in the aftermath of its evident malfunction. In Section 4, I defend the content of these measures in greater detail, emphasising the fortificational benefits of integrating the distinct tasks that wrongdoers have in the aftermath of moral failure—apology and compensation—*into* the experience of fortificational punishment, and I showcase recent efforts in the “restorative justice” movement in order to clarify the potential benefits that such approaches offer. In Section 5, I take up the question of fortificational punishment's coercive imposition. Finally, In Section 6, I address the objection that the fortificationist theory of punishment disrespects offenders' status as agents, and I explain why it is misconceived. Because my aim is to understand punishment in the well-ordered society, where people have been civically educated to hold liberal ideals, I shall focus on punishment for crimes traceable to motivational failure, though the implications for crimes traceable to epistemic failure will become clear as I proceed.

§1. The Quest for Unity

It is commonplace that the central confrontation of penal theory has long been the standoff between consequentialist and retributive justifications of punishment. That this confrontation continues to rage in theory and in practice is a testament to the fact that both justifications appear to capture important and deep truths.¹¹ Consequentialist accounts capture the crucial truth that what we take ourselves to be doing when we criminalise certain conduct involves an aspiration to *reduce the quantity of wrongdoing*.¹² Retributive accounts capture the crucial truth that punishment must respect the agency and status of the criminal offender *qua* moral person, and that the punishment of an offender cannot simply involve a forward-looking quest to prevent further harm; additionally, it must somehow involve a backward-looking *reckoning* with the wrong already done. There has long been significant pressure to believe that we must jettison one of these basic approaches—a theory of punishment centred on crime prevention, or one treating offenders as moral persons—in order to be consistent. G. W. F. Hegel famously said of deterrent threats, the favoured tool of crime prevention advocates: “To justify punishment in this way is like raising one’s stick at a dog; it means treating a human being like a dog instead of respecting his honour and freedom.”¹³ More recently, R. A. Duff has argued that a concern with crime prevention fails to respect offenders’

¹¹ I use “truth” in a manner friendly to political liberalism. See Joshua Cohen, “Truth and Public Reason,” *Philosophy & Public Affairs* 37, 1 (2009): 2–42.

¹² Note that consequentialist accounts of punishment need not be justified within a broader consequentialist moral theory. One of the peculiar mistakes in penal philosophy is the assumption that the aim of crime prevention can only be justified within a utilitarian framework—as though no institutions or practices for deontological or contractualist theories were ever justified on the basis of their instrumental value. This point is argued effectively by Victor Tadros, *The Ends of Harm: The Moral Foundations of Criminal Law* (Oxford University Press, 2011), p. 22.

¹³ G. W. F. Hegel, *The Philosophy of Right*, translated by T. Knox (Oxford: Oxford University Press, 1942), p. 246.

capacity to see moral reasons and act from them.¹⁴ Many protest in response that a system of punishment that is not designed for the productive purpose of reducing crime is not one worth funding.¹⁵

Such hesitation to jettison *either* the idea of crime prevention or the idea of treating offenders as moral persons has led to a flourishing array of so-called “mixed theories” of punishment, which attempt to unify the two ideas into a single, coherent view. But often the unity achieved is superficial, as either one aim has been subordinated to the other, or the equal status granted to both remains theoretically unmotivated. John Rawls’s “Two Concepts of Rules,” often described as a mixed theory, explains how retributive and consequentialist insights can rest together by invoking the rule-utilitarian distinction between the justification of a practice and the justification of a rule *within* that practice. But Rawls explicitly says that retributive insights are accommodated only because of their “obvious utilitarian advantages,”¹⁶ precisely the sort of accommodation of which devout adherents of retributivist convictions would disapprove.

B. Sharon Byrd’s influential interpretation of Immanuel Kant’s philosophy of punishment also offers a mixed view. Byrd’s thesis is that Kant implicitly relies upon a distinction between the justification of the *threat* of punishment and the justification of the *execution* of punishment. When the state *threatens* punishment, its reason is to prevent crime. But when court and penal officials *execute* punishment, their reason is to give people what they deserve. Kant’s most infamous passage on the topic, however, raises trouble for Byrd’s view:

Even if a civil society were dissolved through the unanimous vote of its members (e.g., if an island society decided to dissolve with its members spreading

¹⁴ R.A. Duff, *Punishment, Communication, and Community* (New York: Oxford University Press, 2003), p. 85.

¹⁵ Victor Tadros, *The Ends of Harm*, p. 89.

¹⁶ John Rawls, “Two Concepts of Rules,” *The Philosophical Review* 64, 1 (1955): 3-32, p. 16.

themselves over the rest of the earth), the last murderer within its prison first must be executed so that he experiences what his own deeds are worth and the bloodguilt does not cling to the society that did not insist on this punishment: since, if so, it can be seen as having participated in that public violation of justice.¹⁷

Obviously the execution of the final murderer will not assist in crime prevention in this case. If Kant regards it morally paramount that every person incur “what his own deeds are worth,” why read desert as a *limiting constraint* on punishment, as Byrd does, rather than its proper end?¹⁸ Byrd valiantly endeavours to find unity in Kant’s view despite the apparent tension between his comments on deterrence and the aforementioned insistence on giving people what they deserve. But she succeeds only by subordinating the role of desert in the final analysis.¹⁹

H. L. A. Hart takes up a defence of limiting retributivism, according to which punishment’s aim is consequentialist whereas its limits are retributive. His discussion implies three ways a mixed theory could be structured, and each is either inadequate or underspecified. On the first, the mixed account amounts to cheating: taking one’s favourite parts of discrete, incompatible moral theories and gluing them together. If we think of consequentialist justifications for punishment as originating within a broader consequentialist moral doctrine such as act utilitarianism (as suggested by Hart’s idea that consequentialist justifications could conceivably countenance the imprisonment of innocents), and retributive justifications as originating within some broad deontological or contractualist doctrine such as Kant’s ethics, then a mixed account of punishment that tosses together elements of each worked-out doctrine would have a theoretically

¹⁷ Kant, “The Metaphysics of Morals,” p. 474.

¹⁸ Byrd, “Kant’s Theory of Punishment,” p. 153.

¹⁹ Specifically, Byrd relies on the thesis that public officials, as a matter of “honour,” must discharge all their moral duties prior to undertaking society’s dissolution—even those duties that, in virtue of its prospective dissolution, no longer have a point. Giving people what they deserve is not actually important for Kant’s view. For agreement on that last point, see Thomas E. Hill, Jr., “Kant on Wrongdoing, Desert, and Punishment,” *Law and Philosophy* 18 (1999): 407-441.

unsatisfying character. This implausibility invites us to pursue a second reading of what a mixed account entails. On the second reading, the idea of a mixed account is the necessary upshot of the truth of *value pluralism*: given the impossibility of co-satisfying the moral demands raised by discrete and incompatible values, we must settle for some rough, intuitive compromise. This reading is suggested when Hart calls the best account of punishment “a compromise between distinct and partly conflicting principles” and when he says that “our main social institutions always possess a plurality of features which can only be understood as a compromise between partly discrepant principles.”²⁰ While this reading is more plausible than the first, it, too, has its problems. Crucially, it is not clear that a mixed account whose content is the product of the intuitive, *ad hoc* assignment of differential weights to different, incommensurable values could reasonably be called a mixed *theory*. But Hart elsewhere appears to resist this second interpretation, writing: “what is most needed is *not* the simple admission that instead of a single value or aim...a plurality of different values and aims should be given as a conjunctive answer to some *single* question concerning the justification of punishment.” Rather, he continues, inviting a third reading: “What is needed is the realization that different principles...are relevant at different points in any morally acceptable account of punishment.”²¹ This third reading is the most promising, and it is suggested by Hart’s resistance to Hegel’s aforementioned warning:

it is perfectly consistent to assert *both* that the General Justifying Aim of the practice of punishment is its beneficial consequences *and* that the pursuit of this General Aim should be qualified or restricted out of deference to principles of Distribution which require that punishment should be only of an offender for an offence [i.e., that prohibit the punishment of the innocent].²²

²⁰ H. L. A. Hart, “Prolegomenon to the Principles of Punishment,” in *Punishment and Responsibility*, edited by John Gardner (New York: Oxford University Press, 2008), pp. 53, 62.

²¹ *Ibid.*, p. 3.

²² *Ibid.*, p. 8.

This appeal to consistency avoids the notion of *compromise* found in the second reading. But it, too, is inadequate for my purposes. His discussion of how utilitarianism could play the central justifying aim of punishment, and yet be constrained by retributive principles of justice that render it wrongful to punish innocents, fits under the *first* reading of the mixed account, which simply steals and glues together intuitively promising parts of different established moral philosophies. We thus remain without an account of punishment that explains the *moral connection* between the aspiration to prevent crime and the aspiration to treat every individual offender as an agent.²³

This moral connection matters because we want to know whether and how crime prevention could be pursued in a manner consistent with the political liberal conception of moral personality—our distinctively Rawlsian interpretation of what it is to treat people as agents. In what follows I develop this connection by reconceiving crime prevention—not as some mission by social engineers to reduce society’s crime rates—but as a personal moral objective belonging to every citizen, each aware of her own fallibility and thus properly moved by an aspiration to take measures to reduce the likelihood that she will choose to act wrongly. By showing how a free and equal citizen who commits a criminal offence properly judges herself to have a *duty* to submit to “fortificational punishments,” the consequentialist concern for crime prevention and the retributive injunction to respect criminal offenders as agents are not merely *mixed*, but powerfully synthesised.²⁴ In undertaking this synthesis, I will expose the lie in the

²³ For other criticism of mixed theories, see Richard L. Lippke, “Mixed Theories of Punishment and Mixed Offenders: Some Unresolved Tensions,” *Southern Journal of Philosophy* 44, 2 (2006): 273-95.

²⁴ The idea that criminal offenders are under such a duty is incurring increased popularity. See R. A. Duff, *Punishment, Communication, and Community*, and Victor Tadros, *The Ends of Harm*. Duff discusses the importance of this approach in “Punishment and the Duties of Offenders,” *Law and Philosophy* 32 (2013): 109-127, p. 110. While Tadros calls his own distinctive theory of punishment “The Duty View,” the proposal that offenders are under duties to submit to

popular retributivist refrain that any theory that views crime as the product of causal forces is thereby disrespectful to criminal offenders as agents. For, if agents are *themselves* responsible for managing the forces that threaten to interfere in the successful operation of their own reason, our insistence that criminal offenders undertake such proactive fortification is precisely what treating offenders as responsible agents involves.

§2. The Phenomenology of Criminal Failure

You are a citizen of a well-ordered liberal democracy. Your social relations with your fellow citizens are thus fundamentally just. You and your fellow citizens' mutual commitments are continually fortified by the presence of various fortificational measures, and you have succeeded in realising your vision of a social world in which you live together as free and equal. But one day, something terrible happens. You find yourself moved to violate someone's rights. Maybe you had been tempted to do it for some time. Maybe it occurred to you in the heat of the moment. You knew what you were doing was wrongful. Perhaps you felt the pull of your sense of justice, enjoining you to stop. Perhaps you did not. Whatever the case, the right has been violated. Your fellow citizen's prized possession has been deliberately destroyed. Or she has suffered grievous bodily harm. Or she is dead, by your hands.

How should you feel? What should you do? Firstly, you ought to *recognise* the wrong you have done. It is intrinsically valuable when people expressively affirm the value of others by sincerely avowing their equal moral status. To see that wrongdoers ought to do this, we need only consider our reactive attitudes toward them were they to *refuse* to do this—to *deny* that those they have wronged have the requisite moral standing

punishment is certainly not original; see Plato, "Apology" and "Crito," in *Five Dialogues*, translated by G. M. A. Grube (Indianapolis: Hackett Publishing Company, 2002).

to complain, or to remain silent when asked whether they perpetrated a crime. By recognising the wrong you have done, you honour the fundamental ideal of justice—to regard others as free equals—and affirm your commitment to it. The moral reasons to recognise the wrong done help explain one important reason for the criminal trial—and why one should turn oneself in were citizens to organise such a trial: it is appropriate that those who acted wrongly claim responsibility for their actions.²⁵

Secondly, you ought to *apologise* to those you have wronged. This is naturally related to recognition of the wrong done; it presupposes recognition, but its meaning is not exhausted by that presupposition. When people say, “I know what I did was wrong, but I’d do it again,” they betray a flagrant disregard for the *regulative primacy* of moral norms: the fact that moral norms do not just outweigh, but rather pre-empt, all other reasons for action. No quantity of prudential reasons to kill an innocent person for professional gain could render prudence regulative over morality in that case. For someone to be apologetic for what she has done, she ought to exhibit what I shall call *decisive regret*—not merely *pro tanto* regret—for having performed the wrongful action, and she must do so for the right reasons. And she must *communicate* her recognition and regret to the appropriate audience: her victims and the wider political community. Victims of crimes appropriately feel slighted, even outraged, by having been wronged, and criminal offenders have no complaint when they receive censure by their victims and fellow citizens.

Thirdly, if possible in the case, you ought to offer *compensation* for the crime perpetrated. For most economic crimes, compensation is relatively straightforward: the

²⁵ This claim has been defended in T. M. Scanlon, “Giving Desert Its Due,” *Philosophical Explorations* (forthcoming). For scepticism about the alleged importance that a wrongdoer feel guilty, see Gilbert Harman, “Guilt-Free Morality,” in *Oxford Studies in Metaethics*, Volume 4, edited by Russ Schafer-Landau (Oxford: Oxford University Press, 2009).

offender ought to compensate the victim for the money stolen, or the value of the burgled goods, as well as the loss to productivity or inconveniences incurred in virtue of their absence. For some economic crimes—stealing, melting down, and selling the gold pendant someone inherited from her grandmother—the value of the stolen good cannot be compensated entirely, for some of its value was non-monetary. As for non-economic crimes, the idea of compensation becomes far more slippery. Rapists should pay for their victims’ therapy. But it is important not to make the crucial mistake that such compensation is somehow compensation *for rape*; it is compensation for the cost of the therapy, the need for which was generated foreseeably by the perpetrator’s act of wrongdoing. *Rape* cannot be compensated for. Consider, too, the example of a murdered child. Suppose the parents win a lawsuit, receiving millions of dollars in the aftermath of their child’s murder. The parents would be seriously mistaken if they believed that a simple bank transfer could “make up” for the loss of their child. As Daniel McDermott writes, “Making this kind of a payment would be like a father who had a duty to provide his child with a certain amount of love attempting to satisfy his duty by handing the child \$100.”²⁶ For liberals, *nothing* can make up for the loss of human life. We should not expect compensation, in other words, to be possible in all cases.²⁷

²⁶ Daniel McDermott, “The Permissibility of Punishment,” *Law and Philosophy* 20, 4 (2001): 403–432, p. 414.

²⁷ This claim is controversial. Victor Tadros attempts to build a theory of punishment out of the intuition that offenders *owe something* to their victim. See *The Ends of Harm*, pp. 273ff. I share the intuition that if I have done something bad to somebody and find myself with an opportunity to help him out at a later occasion, I ought to do it. For scepticism about whether Tadros’s view can ground a theory of punishment, see Duff, “Punishment and the Duties of Offenders,” especially pp. 121ff. Tadros’s theory is rich and sophisticated and would require many pages to discuss that would take me too far afield. For reasons that should already be clear but to which I shall return, I believe theories like Tadros’s—which see penal systems as systems of deterrence—would, when institutionalised, exert a disfiguring effect on the quality of our relations in an ideal well-ordered society. Thus even if Tadros’s vision could be plausible in a non-ideal context—in which citizens have reason to be continually afraid that they are at risk of prospective attack, and therefore would have reason to insist that their previous attackers submit to punishment to deter fresh

Fourthly, beyond compensation, if possible you should *disgorge* the profits of your wrongdoing. This idea reflects the well-worn principle, entrenched in many legal systems, that offenders ought not to benefit personally from their wrongdoing. Suppose Alan steals money from Sandra and invests it to make a profit. In such a case, Alan ought not only return the quantum of stolen money (alongside interest, etc.) to Sandra; he should relinquish the amount generated through investment to the political community, given its proper role as the acquirer of un-owned goods. Suppose that a Wall Street hedge fund skirts some important legal regulation and makes a quick billion. It lacks legitimate claim to the legal title of that wealth, as it has been secured through injustice. It must disgorge it.²⁸

I have identified four courses of action, and the proper attitudes that underlie them, that an offender ought to pursue, and experience, in the aftermath of her moral failure. I have not supplied arguments for them, but simply noted their plausibility in accordance with our considered convictions. By pursuing a deliberately Cartesian approach, I have tried, however artificially, to keep separate our distinct considered convictions about what wrongdoers ought to do in the aftermath of their moral failure. Obviously one thing that many believe to be the *most* important course of action to which offenders must submit in the aftermath of their moral failure—punishment—has not yet arisen, at least not obviously. Given most people’s intuition that it is the most *obvious* thing that a criminal offender has moral reason to submit to in the aftermath of moral failure, that is a curious. Why have we not mentioned it yet?

attackers—the permanent fear of the constant threat of attack that Tadros’s view assumes that citizens experience does not belong to our vision of a well-ordered liberal democracy.

²⁸ See Robert E. Goodin, “Disgorging the Fruits of Historical Wrongdoing,” *American Political Science Review* (forthcoming).

There are three possibilities I shall consider: either punishment is somehow *already* accounted for in the list we have developed; or our list is exhaustive and thus there is no place for punishment; or our list is *not* exhaustive and what remains indeed includes something that can plausibly be described as punishment. I shall endorse the third option, but only after dealing with the first two.

Perhaps punishment has already entered the picture through our consideration of compensation and disgorgement. According to the doctrine of *fair-play retributivism*, punishment just *is* the removal of the unfair advantage that one has secured by breaking the law—restoring the distribution of benefits and burdens in society to its level prior to the crime, which is precisely what the demands of compensation and disgorgement facilitate.²⁹ There are, however, problems with this view. Firstly, consider the observation that many people would hesitate to call these *punishments*. Consider the conceptual formula of what punishment *is*: the imposition of a burden in response to the contravention of a prohibited norm. But the most plausible baseline with reference to which setbacks to interests are measured is a *pre-crime* baseline. To be sure, this cannot be settled until we have a particular *conception* of punishment—a normative penal theory. But the intuition that compensation or disgorgement should not qualify as punishments (or, anyway, as *proper* punishments) is widespread. It does not *seem* right, to use a different example, to say that someone's proper punishment for stealing Betty's television is that he has to wander by Betty's house and give the television back to Betty. *That* was already

²⁹ "Justice—that is punishing such individuals—restores the equilibrium of benefits and burdens by taking from the individual what he owes, that is, exacting the debt." Herbert Morris, "Persons and Punishment," *The Monist* 52, 4 (1968): 475-501, p. 478. See also Dagger, "Playing Fair with Punishment," pp. 473-488. For Dagger's response to the standard objections to fair-play retributivism, which I cannot discuss here, see "Punishment as Fair Play," *Res Publica* 14, 4 (2008): pp. 259-275.

a given. Punishment, it may seem, demands something *distinct* from this.³⁰ Secondly, and more importantly, it is perverse to think that crimes like assault or murder are examples of a citizen unfairly claiming some benefit for herself that the rest of us have abstemiously denied ourselves. From the point of view of our public philosophy—regardless of what reasonable conception of substantive justice, and thus reasonable account of human interests (e.g., primary goods, capabilities, etc.), that one holds—it is implausible that the interest, say, in the pleasure someone experiences in murdering someone can be certified by the political association as a *good* (in the way that the extra income that one incurs from cheating on taxes can be).³¹

So, perhaps punishment enters the picture through the morally important aims of recognising wrongdoing and apologising to victims. R. A. Duff's influential communicative theory of punishment argues as much. According to Duff's view, recognition and apology are so important that offenders' interests ought to be frustrated in order to *induce* them to realise that they have acted wrongly, prompting the appropriate requisite apology:

[T]he proper punitive purpose of hard treatment...would not be so well served by purely formal convictions or symbolic punishments. It serves in part to make our communicative endeavour *more effective*, forcing the criminal's attention onto the implications of his crime; he cannot ignore it as easily as he can ignore or forget a conviction or a purely symbolic punishment.³²

On Duff's view, setbacks to interests are *contingently* necessary for the achievement of recognition and apology, since setbacks such as imprisonment are instrumentally useful

³⁰ The intuition that punishment is importantly different than compensation is captured, perhaps imperfectly, by the distinction between criminal law and tort law. Cf. Duff, *Punishment, Communication, and Community*, p. 161.

³¹ For an argument that punishment not plausibly conceived as a matter of paying a debt, see Daniel McDermott, "Debts to Society," *The Journal of Political Philosophy* 10, 4 (2002): 439-64. For the argument that it *is* plausible to describe murderers as having incurred an unfair benefit, see George Sher, *Desert* (Princeton: Princeton University Press, 1989), pp. 82ff.

³² Duff, *Trials and Punishments* (Cambridge: Cambridge University Press, 1991), pp. 260-61.

in *focusing the mind*. This is an intriguing argument, but it appears susceptible to a crucial vulnerability: what happens in the cases of offenders who *do* recognise the wrongness of what they have done? Here Duff invokes a *distinct* argument: that the setbacks to interest are not merely *instrumentally* useful in assisting repentance; they are *intrinsically* appropriate:

But it should serve too—if it is successful—as a penance which the criminal comes to will for himself: the pain or suffering which begins as a coercive attempt to attract and direct the unrepentant criminal’s attention should become the penitential pain which the repentant criminal accepts for himself.³³

Thus it is *part* of what it *is* to recognise one’s wrongdoing and to apologise that one should undertake a penance. “Repentance,” Duff writes in later work, “is necessarily painful, since it must pain me to recognise and admit (to myself and to others) the wrong I have done.”³⁴ But here Duff is not simply referring to the pains of remorse. He continues:

Repentance, at least with serious wrongs, cannot of its nature be something that is achieved and finished with in a moment, however intense the feelings of guilt might be in that moment. It must go deep with the wrongdoer and must therefore occupy his attention, his thoughts, his emotions, for some considerable time. Someone who claims to have been deeply grieved by the death of a loved one, but who resumes her normal life, thoughts, and feelings after a day, thus belies her claim. Similarly, someone who claims to have repented the serious wrong he did, but who rapidly resumes his normal life and gives that wrong no further thought, thus belies his own claim.³⁵

Duff certainly captures the phenomenology well. Those in a state of mourning or remorse *do* tend to retreat from certain pleasant activities in which they might otherwise engage. Recall the scene in Camus’s *L’Etranger*, in which Meursault frolics jovially on the beach with Marie, who subsequently recoils in dismay and alienation from the jubilant Meursault upon learning that he has recently come from his mother’s funeral.³⁶ I do not dispute this phenomenology, and I agree that my relations would be impaired with

³³ *Ibid.*

³⁴ Duff, *Punishment, Communication, and Community*, p. 107.

³⁵ *Ibid.*, p. 108.

³⁶ Albert Camus, *The Outsider*, translated by Joseph Laredo (London: Penguin Books, 2000).

someone who *did* not experience full remorse—hence why I have endorsed the existence of moral requirements to apologise and recognise. Nevertheless, however likely it is that remorse triggers certain patterns of reclusive behaviour—drawing away from the world, foregoing pleasant activities—three serious reservations remain about whether we should *impose* such reclusive behaviour on people as a punishment. Firstly, because the reclusiveness would be *imposed*, then it might lack the value it would otherwise have. Especially since the reclusiveness is understood as the *consequence* of genuine regret, what value *could* it have without the regret? Secondly, however important it is that wrongdoers apologise and recognise for the wrong done, it is not clear this could justify the elaborate infrastructure of modern punishment.³⁷ Thirdly, there are reservations about whether the reclusiveness really is valuable from the point of view of *justice*. What is the point of someone already racked with guilt to undergo even further painful experiences? Consider the griefless Meursault: if he had been sitting on the sand, crying and expressing to Marie how much he misses his mother—would we still insist that he ought not to have been at the beach that day, no matter what? Taken together, these three worries are sufficient grounds to doubt the idea that the justification for criminal punishment, as we understand it, is its role in either getting people to say, or enabling them to express, how sorry they are.

I have argued that the idea of punishment—the infliction of some setbacks to interests (such as diminutions in freedom) in response to one's wrongdoing—cannot be adequately explained by looking to the morally appropriate aims that offenders have of recognition, apology, compensation, or disgorgement. This brings us to the second possibility: that we have identified an exhaustive list of what an agent in the aftermath of

³⁷ Tadros, *The Ends of Harm*, p. 89. I believe it could nevertheless justify the elaborate infrastructure of the criminal trial, given the significance of affirming the moral status of the victim and providing a forum for censure and apology.

her moral failure should feel and do, and so “penal abolitionism” follows for the well-ordered liberal democracy—and thus should become a goal of our own public policy, here and now. This is a serious challenge.³⁸ Penal abolitionism has at least four powerful arguments going for it. Firstly, punishment is expensive; the resources could be spent elsewhere. Secondly, setting back the interests of offenders in the manner demanded by punishment is bad for those offenders. It cannot be intrinsically valuable for offenders to have their relationships and projects upended in the way submission to punishment familiarly involves—they do not “deserve” to suffer.³⁹ Thirdly, institutions of criminal punishment are dangerous in virtue of the risks they pose of wrongful conviction. Juries are an instance of what Rawls calls “imperfect procedural justice,” not perfect procedural justice.⁴⁰ They do not convict the guilty and acquit the innocent with 100% accuracy. Even if they are the best way to identify wrongdoers, for the foreseeable future their imperfect record will remain. Without wishing to address the thorny matter of what the rate of wrongful conviction would have to be to render a system of punishment *too* dangerous—or whether that is even the right question⁴¹—if there is not a weighty

³⁸ Deirdre Golash, *The Case Against Punishment: Retribution, Crime Prevention, and the Law* (New York: New York University Press, 2005). I thank Joanna Firth for directing me to this reference.

³⁹ The idea that it is intrinsically valuable when wrongdoers suffer is the central thesis of desert-retributivism. Desert-retributivism—the idea that it is simply valuable in itself when those who have caused suffering in others experience suffering themselves, because they deserve it—has been attacked by many thinkers, and I have no intention of contributing to that career-worthy effort in a footnote. As H. L. A. Hart says, desert-retributivism “appears to be mysterious piece of moral alchemy in which the combination of the two evils of moral wickedness and suffering are transmuted into good...” Hart, “Postscript: Responsibility and Retribution,” in *Punishment and Responsibility*, pp. 234-35. For another criticism of this view, which relies on controversial metaphysical theses I am reluctant to endorse, see Derek Parfit, *On What Matters*, Volume 1 (New York: Oxford University Press, 2011), pp. 263-72. See also Gaus, *The Order of Public Reason*, pp. 482ff, and Tadros, *The Ends of Harm*, pp. 73-78.

⁴⁰ Rawls, *A Theory of Justice*, p. 74.

⁴¹ It might be supposed that for contractualists *numbers* should play no role in determining whether an institution like criminal punishment is justified; so long as the weight of *one* person’s objection against the system is stronger than any other individual person’s objection to *eliminating* the system, this should be sufficient. On this issue, see the classic discussion in John M. Taurek, “Should the Numbers Count?” *Philosophy & Public Affairs* 6, 4 (1977): 293-316. Applying the issue

justification for subjecting people to these risks, the system cannot be justified. Finally, institutions of criminal punishment are dangerous in virtue of the potential for wrongful criminalisation. If the democratic legislature issues a morally mistaken directive, many people's lives will be burdened unjustifiably. Unless there are sound arguments for instituting the system, this is yet another argument in favour of stopping my list of duties where I ended it.

Is punishment an indefensible ideal? Consider, again, our archetypal criminal wrongdoer. She has recognised the wrong she has done, incurred the state's condemnation, apologised and expressed sincere regret to her victims, paid compensation, and disgorged the benefits of his crime. Our question has been: *what else should she feel or do?* From our discussions of apology and recognition, it is already clear that she should be frustrated with herself; she ought to have decisive regret for what she has done, and have it for the right reasons. But we can distinguish between two kinds of frustration. *Lamentational frustration* transpires when the offender expresses outrage with herself, regretting what has happened and cursing herself for permitting it to happen. I do not deny that this can be fully appropriate. But the second kind of frustration—*productive frustration*—is the attitude an offender has when she realises she has failed and asks herself: *How did this come to be? How did I let this happen? What could I have done differently? What was going through my mind? What were the steps that led up to this? What step was the pivotal step—was there one? What distractions or temptations made me fail to see the relevant facts, or the reasons for action they gave?* These questions are productive—potentially—precisely because they are not rhetorical. Our idealised offender really asks them and wants to know their answers. She wants to understand, causally, how she came to choose to act wrongly.

to Scanlon, Elizabeth Ashford, "The Demandingness of Scanlon's Contractualism," *Ethics* 113, 2 (2003): 273-302.

I believe that someone who does not undertake this sort of reflective process in the aftermath of moral failure is not conducting herself as an ideal citizen ought to conduct herself. We would be alienated from an agent of this kind. For, even if she apologised to her victims; even if she claimed responsibility and recognised the wrong she had done; even if she had offered compensation or disgorgement—if she did not engage in this sort of process, there would be something amiss with her as a moral agent. This insight adduces pressure to think that our list is not yet exhaustive—there is something more that an agent ought to do in the aftermath of her moral failure, and it has to do with reckoning *intelligently* with the crime itself, understanding how and why it happened—not merely for the intrinsically valuable activity of focusing attention on the action she regrets, but for the constructive purpose of *reducing the likelihood that she will choose to perform the wrongful action again*. A course of action oriented to the achievement of that aim, I shall now argue, is the *fifth* course of action that a criminal offender ought to pursue in the aftermath of moral failure.

If you agree with me that this fifth course of action fits our phenomenology—if you agree that such an agent *would* have moral reason to undertake this course of action—then you are on your way to endorsing what I shall call *the fortificationist theory of punishment*. The next step in the process of arriving at such a theory is to convince you that the moral reason to perform this course of action is weighty enough to qualify as a *duty*. It is to this task that I now turn.

§3. The Moral Duty to Submit to Fortificational Punishment

Why should we think a perpetrator of a given crime has not simply a moral reason, but a *duty*, to pursue courses of action (whatever they might be) that decrease the likelihood

that she perpetrate that crime, or relevantly similar crimes, again? Recall my example from Chapter Two:

Tumour 1. The doctor has told you that you have a benign tumour in your brain that exerts an effect on your practical reasoning, influencing you to choose to act wrongly when you perceive that doing so will serve your narrow, immediate self-interest. The tumour was not caused by any choice or series of choices for which you are responsible. It is inoperable—it is here to stay and is therefore a permanent feature of your mental constitution. However, you can mitigate its effects through various exercises that strengthen the countervailing inhibitions against acting wrongly.

I argued that the intuition that we ought to undertake these exercises rests in reflective equilibrium with our considered convictions, both about the importance of not risking others' well-being, and about the criteria of responsibility. Just as most people would have the intuition that they owe it to their fellow citizens to undertake these exercises, I argued that we ought to respond to the fact of our fallibility in a similar fashion: undertake steps to mitigate the risks it poses to others. And I argued that if we could do so in a manner that *fortified*, rather than simply bypassed or eliminated our agency (as by programming some *DRONES*, or simply committing suicide, would respectively involve), we could sustain the moral quality of our just social relationships. With fortificational winds in our moral sails, we would be more likely to realise our aspirations for a social world in which we freely give justice to one another.

When someone violates a weighty perfect duty by committing a serious crime, her confidence in the likelihood that she will choose to act enduringly as justice requires—even with the assistance of general fortificational practices—is properly shaken. She ought to worry: *If I permitted myself to choose injustice this time, will I permit myself to choose injustice again? Can I be relied upon by my fellow citizens, as a partner in liberal-democratic civic life, to live up to the standards of our civic relationship?* The same intuitive pressure as we saw in

initially in the *Tumour* example returns with raging force. If there ever were evidence of fallibility, it is someone's decision to contravene a legitimate law.

As we saw in the last section, the moral reasons a criminal offender has to undertake additional measures to fortify her sense of justice in the aftermath of its evident failure are weighty. It would be outrageous to suggest that morality *disallows* her from undertaking the measures. But does she have a *duty*? If sufficiently weighty countervailing reasons can be identified, it is plausible to conclude that undertaking the measures is simply morally optional, as opposed to morally required. If not, we can conclude that she is under a duty. What might the countervailing reasons be? That is, what *objection* could an offender have to this scheme?

In a modified version of an objection considered in Chapter Two, our offender might object that she did not *ask* to have an especially severe degree of motivational fallibility. It is therefore *unfair* of her to have to undertake extra measures. So, instead of undertaking the measures, it is instead *her fellow citizens* who should have to undertake extra burdens—perhaps they are tasked with staying out of the offender's way, so as not to risk having their rights violated. Perhaps the person who realises white supremacy is wrong but claims not to be able to help herself from attacking racial minorities, then, can legitimately insist that racial minorities steer clear of the bars she frequents.⁴² But we find this response implausible because the offender *retains* the capacity to choose to do what is right; she continues to conceive of herself as an agent. If she harms others, she is *choosing* to harm them. She experiences her behaviour as the product of her choices—her arms do not whip around slamming into people like a rogue machine, her bloodshot eyes playing catch-up, following her limbs as they wreak a moral havoc she insists, even in the heat of the moment, she did not intend. To be sure, she experiences temptations to act

⁴² R. A. Duff and S. E. Marshall, "Benefits, Burdens and Responsibilities," p. 27.

wrongly, maybe even *burning* temptations. But these temptations make a difference to what she is *likely* to do; they do not make a difference as to what she *can* do. Justice enjoins people to work hard to live up to certain standards. Some people need to work harder than others—and that is simply how the cards fall.⁴³ If an agent properly sees herself as subject to the demands of justice, no other conclusion conceivably follows; “their nature is their misfortune”⁴⁴ with which they must reckon.⁴⁵ Moreover, citizens have a duty to avail themselves of fortificational measures in the general course of moral life precisely so they *don’t* end up acting wrongly. Criminals necessarily must see themselves as having *failed* to accomplish the requisite preparatory work.

I have motivated a fifth duty to which an agent is subject in the aftermath of moral failure: to work to identify the cause of the first moral power’s failure and pursue measures to reduce the likelihood of such a failure’s recurrence in the future. Given that human beings are embodied, fallible entities with limited powers, this process of moral fortification is a process of hard work: of reflecting on the specific law one has broken to solidify one’s understanding of its demands and moral legitimacy; in undertaking introspection as to why one acted in contravention of it; and in resolving intelligently on certain courses of thought and action to decrease the likelihood that one will engage in criminal wrongdoing again. Unlike the sentiment of regret for a murder—something an agent can feel instantly and that (appropriately) haunts him even as he resumes normal

⁴³ In accepting this view, I implicitly accept that moral luck is an acceptable element of our moral understanding. I cannot discuss this vexed topic further here. For the classic discussions, see Thomas Nagel, “Moral Luck,” in *Mortal Questions* (New York: Cambridge University Press, 1979), and Bernard Williams, *Moral Luck* (Cambridge: Cambridge University Press, 1981).

⁴⁴ Rawls, *A Theory of Justice*, p. 504.

⁴⁵ Specifying the conditions under which someone is *mistaken* to see herself as substantively responsible for living up to the demands of justice is a vexed and difficult matter. Where do we draw the line between genuine pathologies of the mind, and simple worse-than-statistically-average fallibility? These are important issues, confirming that one’s self-understanding as an agent is a necessary but not sufficient condition for *being* an agent. But we do not need to have a complete view of how to think about these vexed issues in order for us to construct a general framework.

life—these pursuits really do take time out of one’s schedule and thereby reduce one’s opportunities for pursuing plans, projects, and relationships during the time they demand. In cases of morally egregious crimes in which an agent has demonstrated a quite serious failure to exercise her capacity for justice, the degree of fortification required—and thus the appropriate amount of time and energy to dedicate to it—can be daunting. That it is appropriate to enforce these assignments—I matter to which I shall turn in the penultimate section—makes them satisfy our conceptual formula for what punishment *is*: an imposed burden in response to the contravention of a prohibited norm.

Before proceeding, I want to make three clarificatory comments to stave off objections. Firstly, given what I said in Chapter Two about the criteria for testing the status of a fortificational proposal as reasonable, fortificational punishments—like the content of any fortificational measure—must be proportionately burdensome. That is, a given measure must stand a reasonable chance of accomplishing its objective, and it must be the least burdensome way of endeavouring to do so. Now, it is a compelling assumption that morally graver crimes reflect the presence of greater interferences in one’s successful deployment of the first moral power. Someone who commits a minor infraction—stealing a pack of chewing gum—does not thereby display a seriously malfunctioning sense of justice. It would be disproportionately burdensome to assign her extraordinarily demanding measures—say, those assigned to murderers—given the minor degree of malfunction displayed. To insist that she undertake greater measures than those that stand a reasonable chance of supplying the aspired fortificational assistance would run afoul of the test of proportionate burdensomeness. Now, how to apply this test is in an extremely complex philosophical and empirical question that I postpone for another day. It is important to see that the fortificationist theory retains the retributivist’s insistence that harsher crimes ought to receive harsher punishments, but

not because it is intrinsically fit or appropriate that this be so, but simply because harsher crimes indicate the appropriateness of greater fortification. But our *cardinal* determinations of exactly the content and degree of punishments are fated to remain indexed to on-going empirical reflection and evaluation. This generates “empirical snarls,”⁴⁶ to be sure, for the fortificationist theory. But for a theory that aims not simply to condemn human fallibility, but to do something about it, such complexity simply comes with the territory.⁴⁷

Secondly, because commission of criminal wrongdoing serves an *evidentiary* function for the fortificationist theory of punishment, the character of the circumstances in which a given crime is committed will have an impact on our determination of how susceptible to recidivism the offender is. This, however, runs afoul of the idea that people ought to receive the same punishments for the same crime. But it is hard to see the rationale for this objection in a way that does not presuppose the very desert-retributivist, fair-play retributivist, or deterrent theories that a theory of this character aspires to replace. Given that the nature of the fortificationist theory is to enable the criminal offender to reckon with the content of the past crime in the aspiration that she will reduce the likelihood that she will commit that crime or relevantly similar crimes again, the justification of any given punishment is highly sensitive to exact nature of the moral failure in question. Someone who kills someone in the midst of a bar brawl does

⁴⁶ Matthew Kramer, *The Ethics of Capital Punishment* (New York: Oxford University, 2011), p. 181.

⁴⁷ Andrew von Hirsch’s theory aims at deterring people from committing crime, but requires that crimes be ordered in cardinal ranking of moral seriousness since to do otherwise would miscommunicate to criminal offenders the seriousness of their crime. This proposal is vulnerable to numerous objections—for example, that if deterrence really is the objective of the penal system, then disrespectful worries that offenders will misunderstand what they are being blamed for should not figure as limiting factors. At any rate, what von Hirsch helpfully proposes is that we anchor the punishment scale arbitrarily and then lower it slowly is defended as part of a “decremental strategy” as we discover what it takes to deter crime. The fortificationist theory could follow a similar proposal. Andrew von Hirsch, *Censure and Sanctions* (New York: Oxford University Press, 1993), pp. 40ff.

not express the same degree of defectiveness in her command of her first moral power than someone who is a professional hit man. This fact is already accommodated (though its significance is understood differently, depending on the theory in question) through discretionary sentencing schemes, and also by including temporal and circumstantial elements in the specification of different crimes' defining contents.

Thirdly, suppose an offender commits a particular crime that she is, for whatever reason, incapable of perpetrating again. Does it therefore follow that she should not be punished? As Warren Quinn suggests when entertaining a similar objection, "this implication would be absurd."⁴⁸ But why? Suppose I cheat on my taxes a day before the tax system is overhauled from a U.S. to UK model, where my tax obligations are wholly withdrawn from my salary. Does this mean I ought not to be punished? Not at all; punishment remains wholly appropriate, for there remains a wide degree of relevantly similar crimes, in that the same kind of disregard involved in tax-cheating could render one susceptible to temptations to commit other kinds of fraud.⁴⁹ The determination of relevant similarity presupposes some classification of crimes and the distinctive failures that they involve—a worthy topic for a subsequent fortificationist research programme.

Many will be wondering how the fortificationist scheme of punishment differs from prevailing approaches to rehabilitation. These approaches have incurred the ire of retributivists for their apparent failure to regard criminal offenders as agents, viewing them instead as blameless patients who must be pitied and treated. Does the fortificationist theory of punishment avoid these objections? It does. But I plead patience: defeating this final objection is best pursued once my skeletal description of the fortificationist theory of punishment is complete.

⁴⁸ Quinn, "The Right to Threaten," p. 331.

⁴⁹ I thank Victor Tadros for this example.

§4. The Content of Fortificational Punishment

What ought fortificational punishment actually involve? There is no serious controversy about the claim that reflection about reasons for action can make a difference to what people are likely to do. That is simply what practical reasoning *is*. The controversy arises in explaining what transpires during such a process, especially with respect to moral motivation. Moral psychology is a complex business. As with any part of the theory of stability, different theories of moral psychology yield different approaches. As I argued in Chapter Two, what is needed is a family of candidate accounts of how we ought to understand the phenomenon of moral choice and moral failure that can give us traction in specifying determinate proposals for the content of fortificational punishment. In exploratory spirit, without any intention of exhaustiveness and with all deliberate generalisation, consider the basic upshots of four different families of thought in moral psychology, which I mentioned in Chapter Two.

Firstly, according to the Humean approach, the key to moral motivation is the *desire* to be moral: the cognitive judgement that an act is wrong is insufficient to motivate. To see this, we need only imagine a sociopath who seems to *understand* what morality requires, but is left cold motivationally. And when we reflect on what we have that he lacks, we see that it is not that he believes something we do not; it is that we *want* to be moral—we have a *passion* for (to shift to our own, non-Humean terminology) treating one another as free and equal.⁵⁰

⁵⁰ Stefan Kalt, “David Hume on the Motivating Effect of Moral Perception: Internalist or Externalist?”, *History of Philosophy Quarterly* 22, 2 (2005): 143-160; Sharon Krause, “Hume and the (False) Luster of Justice,” *Political Theory* 32, 5 (2004): 628-55; Nancy Schaubert, “Hume on Moral Motivation: It's Almost like Being in Love,” *History of Philosophy Quarterly* 16, 3 (1999): 341-366; Elizabeth S. Radcliffe, “Hume on Motivating Sentiments, the General Point of View, and the Inculcation of ‘Morality’,” *Hume Studies* 20, 1 (1994): 37-58. For David Hume’s account, see *A Treatise of Human Nature*, second edition, edited by L. A. Selby-Bigge and P. H. Niddich (Oxford: Oxford University Press, 1978), pp. 457ff. There is some controversy about whether Hume

Secondly, according to the rival Kantian approach, what motivates an agent to act is her *conviction* that it is the right thing to do. For Kantians, moral success is achieved by *taming* desires and *subordinating* them to one's reason, not (*pace* the Humeans) *through* one's desires.⁵¹ Though Kant's own view is tethered to his own controversial metaphysics, its appeal can also be located within our own moral phenomenology. Especially in our own non-ideal circumstances, in which social arrangements are significantly less than fully just, it is often extremely demanding to do what morality requires. Yet when we ask ourselves *why* we choose to do it, we do not tell ourselves a story about weighing and counting our different desires; we simply say it was right.⁵² Moral convictions *themselves* are what motivate agents to do what is right. Humeans, of course, offer their responses to this objection, which elicit the predictable counter-objections, and the debate continues.⁵³

Thirdly, according to the Aristotelian approach, moral success is achieved because the possession of certain *virtues*—certain habituated temperaments that dispose their holder to act predictably in certain reasonable ways. Aristotle emphasises that virtues, by their nature, are things one takes pleasure in exercising.⁵⁴ While this account of motivation is explicitly conceived as anchored in the moral doctrine of “virtue ethics,”

himself adhered to this view; see, for example, Ingmar Persson, “Hume — Not a ‘Humean’ about Motivation,” *History of Philosophy Quarterly*, 14, 2 (1977): 189–206.

⁵¹ Andrews Reath, “Kant’s Theory of Moral Sensibility,” *Agency and Autonomy in Kant’s Moral Philosophy* (Oxford: Oxford University Press, 2006); Philip Stratton-Lake, *Kant, Duty, and Moral Worth* (London: Routledge, 2000); Josefine Nauckhoff, “Incentives and Interests in Kant’s Moral Psychology,” in *History of Philosophy Quarterly* 20, 1 (2003): 41–60. Michael Weber, “More on the Motive of Duty,” *The Journal of Ethics* 11, 1 (2007): 65–86.

⁵² Russ Shafer-Landau, *Moral Realism: A Defence* (Oxford: Clarendon Press, 2003), p. 123.

⁵³ See, for example, Karl Americks, “Kant, Hume, and the Problem of Moral Motivation,” in *Kant and the Historical Turn: Philosophy as Critical Interpretation* (New York: Oxford University Press, 2006). For an attempted synthesis, see Philip Clark, “Kantian Morals and Humean Motives,” *Philosophy and Phenomenological Research* 68, 1 (2004): 109–126. For a defence of the Kantian approach against some enduring objections, see Richard Galvin, “Does Kant’s Psychology of Morality Need Basic Revision?,” *Mind* 100, 2 (1991): 221–36.

⁵⁴ Aristotle, *Nicomachean Ethics*, edited and translated by Roger Crisp (Cambridge: Cambridge University Press, 2000), p. 14.

we can locate it in the familiar sentiment that moral success flows from someone's possession of good character.⁵⁵ On this view, someone who succeeds morally does so because she has the character traits that dispose her to succeed morally, and has them because she has been habituated to have them.⁵⁶

However, research in social psychology has inspired many philosophers to subject this account to attack—bringing me to the fourth approach. According to this approach, reliable moral success is the result of *contexts* that *elicit* moral conduct. The internal workings of the mind—with its jostling sets of beliefs and desires—is not where the real motivational action lies for human behaviour. Experimental studies demonstrate that seemingly established character traits vanish under pressure. Stanley Milgram's famed experiments, which exposed the tendencies among perfectly ordinary citizens to act wrongly when their context shifted,⁵⁷ have been mustered as evidence by philosophers such as Gilbert Harman and John Doris to dispute the idea that there is such a thing as stable character traits.⁵⁸ The key, according to these *situationist* moral psychologists, is simply to engineer the right situations to bring out the best in people. Virtue ethicists have, of course, responded by suggesting that the situationists have

⁵⁵See especially Michael Slote, "Virtue Ethics and Democratic Values," *Journal of Social Philosophy*, 24, 2 (1993): 5–37, *Morals from Motives* (Oxford: Oxford University Press, 2001), *Moral Sentimentalism* (New York: Oxford University Press, 2010), and *The Impossibility of Perfection: Aristotle, Feminism, and the Complexities of Ethics*, New York: (Oxford University Press, 2011). See also John McDowell, "Virtue and Reason", *Monist* 62 (1979): 331–50; Matt Stichter, "Virtues, Skills, and Right Action", *Ethical Theory and Moral Practice* 14, 1 (2011): 73–86; For some scepticism on whether we ought to speak of "virtue ethics," as a unified category, see Martha Nussbaum, "Virtue Ethics: A Misleading Category?", *The Journal of Ethics* 3, 3 (1999): 163–201.

⁵⁶ Aristotle, *Nicomachean Ethics*, p. 5.

⁵⁷ Stanley Milgram, *Obedience to Authority: An Experimental View* (London: Taylor and Francis, 1974). See also Philip Zimbardo, *The Lucifer Effect: How Good People Turn Evil* (Reading: Ebury Publishing, 2007).

⁵⁸ Harman, "Moral Philosophy Meets Social Psychology," and John M. Doris, *Lack of Character: Personality and Moral Behavior* (Cambridge: Cambridge University Press, 2002).

misunderstood the sort of virtue they have been talking about, and the scholarly debate continues onward.⁵⁹

Each of the four broad approaches I have canvassed has important implications for the theory of punishment. Humean accounts inspire an offender to reflect on her desires, investigating what they are, their role in the process that resulted in her wrongful action, and how she might cultivate new desires that offered her a better sense of congruence between what she takes to be right and what her desired good involves. Kantian accounts inspire offenders to reiterate the reasons they have for standing firm in the face of temptation, scrutinising the forces that may have influenced their wrongdoing to identify their irrationality and resolve not to disgrace one's capacity for reason by acquiescing to them again. Aristotelian accounts inspire offenders to think about what bad habits they have cultivated that may have contributed to their moral failure, and how they can try to cultivate better habits to replace them. And situationist accounts inspire offenders to take seriously what elements of their context exacerbated their pernicious inclinations or weaknesses, reflecting on how they might avoid such contexts in the future or take proactive measures to alter the problematic environments themselves.

Just as the public sphere of a well-ordered liberal democracy ought to be filled with discussions about the nature and functioning of the economy, so too should it be filled with discussions about how we ought to understand the nature and character of moral psychology, with different candidate conceptions of reasonable moral psychology subjected to democratic scrutiny for their scientific plausibility. Each family of theories I have addressed enjoys its fair share of phenomenological and social-scientific

⁵⁹ For attempts to defend virtue ethics against Harman's criticisms, see Joel Kupperman, "Virtue in Virtue Ethics," *The Journal of Ethics* 3, 2/3 (2009): 243-55; Maria Merritt, "Virtue Ethics and Situationist Personality Psychology," *Ethical Theory and Moral Practice* 3, 4 (2000): 365-83; Christian Miller, "Social Psychology and Virtue Ethics," *The Journal of Ethics* 7, 4 (2003): 365-92.

corroboration. Frustratingly, most philosophers who work on moral psychology tend not to be interested in teasing out the implications of their work for their work for normative practice. Their concerns are metaethical: they want to know what the correct structure of moral truth implies for the correct structure of moral motivation, and especially vice-versa; so, J. L. Mackie explores moral psychology to vindicate his famed “error theory,”⁶⁰ Bernard Williams to vindicate reasons internalism,⁶¹ and Gilbert Harman to vindicate his distinctive brand of moral relativism.⁶² However, political liberalism aims to be neutral on such metaethical debates.⁶³ It engages moral psychology so that its adherents may learn how to understand the character of their own moral failure and resolve intelligently not to repeat it.

Given the reasonableness of proposals associated with each, it would be perfectly legitimate for proposals associated only with one or some such views to be democratically enacted. But strikingly, and promisingly, there is substantial *compatibility* among these different approaches I have sketched.⁶⁴ No one strategy precludes the others, when untethered from their respective sectarian foundations (e.g., Kantians and Humeans *both* agree that certain malicious desires can interfere with one’s commitment to justice). This observation helps refute the worry that the fortificationist theory of

⁶⁰ J. L. Mackie, *Ethics: Inventing Right and Wrong* (London: Penguin, 1990).

⁶¹ Williams, “Internal and External Reasons,” in *Moral Luck* (Cambridge: Cambridge University Press, 1981).

⁶² Harman, “What is Moral Relativism?,” in *Explaining Value*, pp. 20ff.

⁶³ Political liberalism is compatible with several, though not all, substantive metaethical positions. It is compatible with moral realism and constructivism. It may even be compatible with certain non-cognitivist doctrines such as emotivism. If someone avows reliably that the world ought to be arranged in accordance with the political liberal conception of moral personality, simply because she experiences a burning, overriding passion for the creation of that world, why insist that she be excluded from the overlapping consensus? This is a difficult question, worth further pursuit.

⁶⁴ See, for example, Steven M. DeLue, “Aristotle, Kant and Rawls on Moral Motivation in a Just Society,” *American Political Science Review* 74, 2 (1980): 385-93; Christine Korsgaard, “From Duty and for the Sake of the Noble: Kant and Aristotle on Morally Good Action,” in *The Constitution of Agency*, pp. 174ff.

punishment is bound to resist consensus, resulting in a system of inconsistent punishments motivated by distinct theories. To complete the refutation of that worry, I shall spend the remainder of this section sketching some concrete examples from actual penal practice of fortificational punishments that rest in concord with some or all of the approaches I have sketched.

Firstly, consider programmes that enjoin offenders to re-enact and discuss elements of their similar crimes to diagnose the anatomy of the actual process of crime commission. For example, the CHANGE Project is a Scottish program in which men convicted of domestic violence attend group seminars in which participants partly re-enact and then discuss scenarios involving spousal abuse. The seminars are emotionally exhausting and intellectually challenging.⁶⁵ The central aspiration of CHANGE is to secure offenders with a clearer understanding of the forces that move them to act wrongly—to identify the particular temptations and external factors that dispose offenders to engage in spousal violence. So described, it rests comfortably with Humean moral psychology—inquiring whether the wrong balance of desires is part of the problem; Kantian moral psychology—affirming that reason did not prevail to mitigate the influence of malicious desires; Aristotelian moral psychology—working to develop strategies for the offender to practice to habituate himself to have certain skills; and situationist moral psychology—working to determine regrettable elements of the external environment that may have disposed one to do what one otherwise may not have done.

Secondly, consider punishments of a more scholarly cast, focusing on understanding *why* a particular crime was wrong. Most moral psychologists endorse the thesis that an improved understanding of why an action is wrong will result in a decreased likelihood of performing the action in the future. While each school has its

⁶⁵ Duff, *Punishment, Communication, and Community*, pp. 102-03, 108-109.

own views on why this is so, the empirical literature suggests it might be effective simply because it makes the offender associate the crime itself more directly with the experiences she has had reflecting about its wrongness and how to avoid doing it again.⁶⁶ (For example, if those who vandalise are systematically likely to associate vandalism with the idea that they are acting wrongly, it may result in a weakened likelihood to engage in it.) The upshot is that *epistemically fortifying* assignments may yield *motivationally* fortifying effects. This is fortuitous because it enables our account of punishment to attend to citizens who did not believe that their actions were wrongful—rare though we hope such persons are in the well-ordered society.⁶⁷ Often these epistemically fortifying punishments are pursued through reading works on philosophy and debating the wrongness of offenders' crimes in the context of great moral and philosophical debates. One such programme in Texas centred on the teachings of Socrates, engaging offenders by “discussing the meaning of concepts such as knowledge, wisdom, ignorance, self-interest, mistake, voluntary, involuntary, happiness, and virtue,” with an eye toward assessing the implications of what they did for their particular crimes, with beneficial effects to show for recidivism rates.⁶⁸ Many such programmes report beneficial effects. Alan Smith ran philosophy classes in English prisons, entreating prisoners to discuss thinkers from Aristotle to Rawls. He even brought in local MPs and engaged them on issues from “the war in Iraq, about freedom and the political machine, about public

⁶⁶ E.T. Higgins, W.S. Rholes, and C.R. Jones, “Category Accessibility and Impression Formation,” *Journal of Experimental Social Psychology* 13 (1977): 141-54.

⁶⁷ For those cases in which punishment will involve engaging the offender's *moral beliefs* for the sake of trying to change them, the fortificationist theory would recommend measures similar to those defended by Jean Hampton, “The Moral Education Theory of Punishment,” *Philosophy & Public Affairs* 13, 3 (1984): 208-38. However, as I have said, this is the rare case, or so I imagine, in a well-ordered liberal democracy with widespread civic education and equality of opportunity. The distinctive scholarly contribution of the fortificationist theory of punishment is its role in *motivational* fortification.

⁶⁸ Richard Layton, “Discussion Group Report: Socrates Changes the Lives of Present-Day Prison Inmates,” http://www.humanistsofutah.org/2000/SocratesChangesLivesPrisonInmates_DiscGrp_08-00.html, accessed on 20 November 2012.

spending and about the necessity of representative democracy.”⁶⁹ At Jessup Correctional Institutional, prisoners read Kierkegaard, Frederick Douglass, and Buddha, writing final essays at the class’s end on competing conceptions of freedom and the ethics of war.⁷⁰ On this view, however, such critical engagement would not be a distracting ancillary activity to the experience of incurring the unfreedom of incarceration.⁷¹ Rather, the whole *point* of the incarceration—of retreating from the life that engendered the temptations to act wrongly in the first place—would be such fortificational pursuits. Incarceration—in those rare cases in which it is preferable to, say, mandatory evening classes—would be conceived as an extended retreat consisting of burdensome fortificational activities. I am inclined to think that the mocking reaction that many would display toward such proposals shows how complacent we have become to the brutal excesses of contemporary criminal punishment.

The third proposed focus of policy consensus is also justified on the basis of the claim that epistemic fortification can lead to motivational fortification. While there are many community sentencing schemes in contemporary penal practice, whereby agents are required to pursue certain socially beneficial projects, a sub-class of them is particularly intriguing: they enjoin work that relates, in some substantive manner, to the wrongness of their crime. Consider vandals who are required to clean up the public square; rapists who are required to participate in seminars teaching others about the

⁶⁹ Alan Smith, “Locke and Key,” *The Guardian*, 20 January 2004, available at <http://www.guardian.co.uk/education/2004/jan/20/furthereducation.uk2>, accessed on 11 November 2012.

⁷⁰ Daniel de Vise, “College student’s philosophy program brings Plato and Buddha to a Md. prison,” *The Washington Post*, 1 September 2011, available at http://articles.washingtonpost.com/2011-09-01/local/35273403_1_maximum-security-prison-paper-clip-inmates, accessed 5 November 2012.

⁷¹ Of course, the prisoners at Jessup were serving a life term—a setback to freedom that cannot be justified on the fortificationist model. Toward *what* are they fortifying themselves? While some mentally disabled or psychotically dangerous offenders may be justifiably imprisoned for life, this is best understood as *justified incapacitation*, not fortificational punishment.

horrors of violence against women; tax cheats required to help prepare materials for distribution to the general public on the damage to the political system resulting from tax fraud; and so on. These are generally understood to be ancillary side-components to the normal burdens of punishment. On a fortificational view, they would *be* the punishments.

The final prospective area for policy consensus among adherents to different moral psychological schools can be gleaned by remembering what it is that criminal offenders have reason to do in the aftermath of their criminal failure *besides* submit to fortificational punishment: recognise the wrongness of what they did, apologise to victims, and pay up compensation and disgorgement when possible. In a well-ordered liberal democracy, I propose that such tasks would be *integrated* into the process of punishment, harnessing them for their epistemically and motivationally fortifying potential. Consider the duties to recognise and apologise. In the case of burglary, this will involve an account of the damage done to the victim's possessions, the frustration of her sense of security, and the disruption of the general scheme of property expectations that grant predictability to people's lives. This process of communication, importantly, can be an extremely emotionally draining experience to endure, since, *as* a communicative process, the offender would be required to listen to the victim's outrage.

The increasingly popular school of "restorative justice" offers numerous examples of programmes that feature dialogue between victims and offenders, in which offenders are invited to communicate their regret and must endure the outrage and condemnation levelled at them by their victims. These restorative activities are often justified within the broader retributivist framework, as victims and offenders work together to determine what quantum of suffering is appropriate, given the crime, but they would not on the fortificationist view. The sessions take many forms. Even if they

do not engage with their particular victim—either because the victims do not want to, or because their victims are dead and their loved ones are unwilling—these programmes often arrange offenders meet with victims of other crimes, serving similar purposes. Offenders must answer such questions as: “What happened?” “What were you thinking of at the time?” “Who do you think has been affected?” “In what way?” “What can you do to make things right?” Victims who participate are asked questions such as: “What did you think when you realized what had happened?” “What impact has this incident had on you and others?” “What has been the hardest thing for you?” “What do you think needs to happen to make things right?”⁷² If Scanlon is right that the nerve of human beings’ sense of justice, no matter what society they live in, is the desire or willingness to justify oneself to others on terms they could not reasonably reject, we should not be surprised by the readiness of people to apologise for what they have done when they confront their victims.

Consider, too, the issue of compensation. Requiring vandals to repair the damaged property themselves may not be sufficient for their fortification, but it can certainly serve as part of it. In some cases, it might be cheaper for all concerned for vandals to issue a monetary compensation for the value of the labour involved in cleaning up the damage they have caused. But there may be important fortificational aims achieved by having *them* do it.⁷³ Even if the work performed by criminal offenders is not *literally* cleaning up the exact mess they made through their actions, assignments involving labourious activities can serve crucial fortificational aims. When Gwen McIvor examined

⁷² Ted Wachtel, “My Three Decades of Using Restorative Practices with Delinquent and At-Risk Youth: Theory, Practice and Research Outcomes,” A paper presented at the First World Congress on Restorative Juvenile Justice, Lima, Peru, 5 November 2009, available at [http://www.unicef.org/tdad/2tedwachtel\(1\).pdf](http://www.unicef.org/tdad/2tedwachtel(1).pdf), accessed 15 November 2012.

⁷³ Caroline Ball, Kevin MacCormac, and Nigel Stone, *Young Offenders: Law, Policy, and Practice* (London: Sweet & Maxwell, 1995), p. 259

Scotland's implementation of community service orders, she found such schemes were reported as valued by offenders when the assignments "maximised their contact with beneficiaries," "enabled offenders to acquire practical or inter-personal skills," and "provided offenders with a sense of benefit to recipients." Those who regarded it as worthwhile (compared to those who claimed it not to be worthwhile) were "slightly less likely to be reconvicted," "reconvicted less often in the three years following sentence," and "markedly less likely to be reconvicted of offences involving dishonesty."⁷⁴

There are obviously many ways in which to interpret what is happening in these situations—why the encounters I have described have led to the reductions in crime recurrence that many have observed. Democratic citizens who are partisans of a particular school of moral psychology will have their preferred description of what is happening. What matters is not that we figure, once and for all, the deep workings of moral psychology. What matters from the perspective of the properly thinking criminal offender is that she try *something* reasonable to fortify her first moral power in the aftermath of its defective operation. In a society in which the fortificationist theory of punishment prevails, we might imagine the judge at a sentencing tribunal informally chastising an offender as follows: "You are under a standing moral requirement, as a fallible person, to keep yourself together, morally speaking. Through your criminal choice, you have demonstrated that you do *not* have it together. *Get it together.*"

§5. The Enforcement of Punishment

⁷⁴ Gill McIvor, *Sentenced to Service: The Operation and Impact of Community Service by Offenders* (Aldershot: Avebury, 1992).

5.1. Punishment and hypothetical self-binding

Fortificational punishments are burdens assigned in response to wrongdoing. But thus far I have been deliberately ambiguous as to what, exactly, this “assignment” involves. Now the importance of submitting to additional fortificational measures in the aftermath of moral failure is *morally required*; it is “assigned” by morality. The particular content of that moral duty is specified through the democratic process, which issues an authoritative directive to submit to punishment of a certain character—and so long as the directive is reasonable, given the fortificational criteria sketched in Chapter Two—it is a morally legitimate legal duty. Thus the assignment we have is both moral and legal. But I have insisted throughout this dissertation that there is no necessary connection between authoritative legal directives and coercion. Thus the sense of “assigned” I have used has not involved the idea that people ought to be *forced* to accept their punishment. Rather, all I have tried to establish is that they have a *duty* to submit to punishment. Does this mean that criminal offenders who refuse to undergo punishment are permitted simply to refuse?

To get at this concern, consider the relation between the system of fortificational punishment and citizens who are not presently subject to its measures. There are two elements of their relation: they are its *democratic designers*, and—as potential offenders—its *potential participants*. Consider the perspective of our partly idealised citizen in the well-ordered society, concerned about the prospect of her own moral failure. She has never broken the law, and she aspires never to do so. But she avails herself considerably of her society’s fortificational measures to maintain the successful operation of her sense of justice. She is thus intensely aware of her own fallibility. She fears she may someday choose to break the law. She is not prepared to join with her fellow citizens to create a system of deterrent threats, because she is aware of the disfiguring effect that would have

on her relationships with them. Yet thinking ahead to the prospect of criminal action, she resolves to herself: “If it turns out to be true that I *am* so fallible as to affront my fellow citizens in such a manner as to engage in serious criminal wrongdoing, it will be of the utmost importance that I submit to additional fortificational measures immediately. I will have, through my criminal choice, identified myself as someone who cannot with confidence be predicted to live up to her obligations as a free and equal citizen in the future. I must restore myself to that standing. Therefore, if for whatever reason the time comes and I *refuse* to undertake the necessary work, let it be so that I am *forced* to submit to the fortificational measures. For, only by undertaking them can I return to my standing as the person whom I properly am committed to be: the sort of person who can be depended upon to freely give justice to others.” Thus our reasonable citizen, aware of her own fallibility, would agree to *bind herself* to undertake fortificational punishment in the eventuality of a choice to engage in criminal conduct.

The theoretical upshot of the contention that citizens have a duty to self-bind in the way I am arguing is two-fold. Firstly, it means that they have a general fortificational duty to create and maintain a system of punishments that is there for them if the time comes that they incur a duty to avail themselves of such systems of punishment, with the power to enforce if necessary. Secondly, it means that they cannot complain when subjected to punishments, for they are under a duty to maintain their standing in the political community as the sort of person who can be relied upon to act justly by her free and equal citizens—precisely the sort of person demanded by the ideal of just civic relationships. A criminal choice is glaring evidence that a person’s sense of justice is

defective and in need of fortification. If she refuses to undertake it, she is failing seriously as a citizen tasked by justice with managing her fallibility.⁷⁵

I want to consider briefly an objection that I have been delaying until now. This objection asks: if punishment of an offender is justified because of the *evidentiary* function of her wrongdoing—the fact that the wrongdoing demonstrates a fact about her that makes the pursuit of additional fortificational measures appropriate—does it not also follow that those who *also* face strong temptations to act wrongly, but for whatever reason have not committed a crime, *also* have a duty to undertake fortificational punishment? The answer is that they certainly have a duty to undertake *fortification*. Someone who experiences violent temptations to lash out at others has a duty to avail herself of fortificational measures that exist for precisely such people. But it will not be *punishment* in such a case, for it would not involve any coercively imposed setback to her interests. The more difficult case is someone who avows that she has temptations to act wrongly but also avows that she does not plan to do anything about. I think, however, there is a strong argument to be made in favour of the position that a person's mere *announcement* that she holds a particular temptation is insufficient evidence to ground the justification of a fortificational imposition. The ideal of freely fortified just social relationships grounds a presumption of faith in human beings' likelihood of engaging in the requisite fortificational measures without having to be forced to do so. As Duff writes, "To treat someone as a responsible fellow citizen is to treat her as

⁷⁵ Some might argue: even if people have a duty to consent to something, their duty is to *consent* to it; thus they cannot be forced. I believe this objection fails to distinguish between those duties whose central aims are the realisation of some state of affairs and duties whose central aims are, for example, the expression of certain forms of respect. This objection holds for the latter, but not for the former—even though, admittedly, forced submission to some state of affairs lacks the intrinsic value that free submission would. This objection is inspired by Mark C. Murphy, "Acceptance of Authority and the Duty to Comply with Just Institutions: A Comment on Waldron," *Philosophy & Public Affairs* 23, 3 (1994): 271-277.

someone...whom we should trust not to commit crimes unless and until she gives us good reason to qualify or suspend that trust.”⁷⁶ This focus on external conduct is not idiosyncratic. One’s external conduct *just is* the reflection of the free operation of one’s sense of justice. Since the role of the sense of justice is to move people to do certain things and refrain from doing other things, if one freely succeeds in doing that, there are no further questions. Criminal conviction is a judgement that the offender has not chosen to comport her conduct with the required standards, and the expression of a justified demand the person work hard to ensure that it does not happen again. But so long as a person works hard to keep herself out of trouble and to keep her temptations at bay—to remain a citizen of full standing in our standard relations, stabilised by fortificational measures—coercively subjecting her to additional measures is unjustified, because it has not been proven appropriate.⁷⁷

5.2. Asymmetric enforcement and the moral quality of social relations

There is a puzzle about my defence of punishment’s coercive enforcement. While duties to submit to punishment are properly enforced, throughout this dissertation I have *resisted* the idea that we ought to conceive of our *fundamental* duties of justice as enforced. But if it is so important that we have institutions standing by to enforce the special

⁷⁶ Duff and Marshall, “Benefits, Burdens, and Responsibility,” p. 29.

⁷⁷ Consider the bizarre case in which God tells everyone that Seth is fated to act wrongly at t2 unless he works hard to fortify himself. Would it be justified to subject Seth to fortificational punishments at t1? This seems to be a problem for my theory, since it strains linguistic usage to call that *punishment*. But the theory has a response. Under such circumstances, Seth himself would have a duty to submit to some fortificational measures at t1, given that God has given him excellent reason to believe he poses a threat. Indeed, given what he knows, he would be acting *wrongly* if he *refused* to undergo such fortificational treatment at t1. Suppose Seth refuses. In such a case, the state cannot rationally sustain its presumption of faith that Seth will not act wrongly at t2; *ex hypothesi*, God has told the state that Seth *will* act wrongly unless he pursues these measures, and Seth refuses to pursue them. In such a case, the state may subject Seth to fortificational punishment *in response to the moral failure exhibited by refusing to fulfill a clear fortificational duty*. It thereby satisfies the conceptual formula of punishment.

fortificational duty to submit to punishment, why is it not so important to have institutions standing by to enforce general duties of justice generally? Given the widespread conviction that it is better that crimes not be committed at all than be committed, this is a curious puzzle indeed.

This solution to the puzzle takes us back to Chapter Two. In that chapter, I considered Rawls's view that stability—the freely willed achievement and maintenance of justice, traceable to the successful operation of the sense of justice in the face of potentially countervailing forces in human nature—is justificatorily significant for our ideal of justice. I argued that it is not. But I also argued that stability nevertheless retains its normative significance, and in two crucial ways. Firstly, understanding how it is that our ideal of justice will be realised in the face of countervailing pressures matters because *justice* matters. We do not take risks with others' rights. Part of what it *is* to have a commitment to justice, I argued, is to think and work hard to ensure that one will live up to its substantive demands—that one will not be subverted by the forces of instability, be they epistemic or motivational, in her attempt to ascertain and do what is right. Secondly, freely willed stability matters because it is a *constitutive* element of fully just social relationships. In order for a relationship to have its full value, its participants must have reason to believe that other participants are engaging in it for the right reasons. The protection of individual rights secured by manipulating or forcing people into “respecting” one another does not belong to our ideal vision of a well-ordered liberal democracy. That ideal does not tolerate the disfigurement of our civic infrastructure of standing threats of punishment issued from the state to its citizens.

This explains the apparent asymmetry between enforcing fundamental duties of justice and enforcing fortificational duties of punishment. It may seem, however, that I have not eliminated the puzzle, only relocated it. For *surely* someone who *refuses* her

punishment is someone with whom I thereby have an impaired relationship. Why not also *insist* that punishment be freely willed? The reply is this: when someone has chosen to commit a criminal act, her relations are *already* impaired by virtue of her decision. Indeed, if she resists the fortificational punishment, they stand to become *even more* impaired. Recall the reasons why citizens are under a duty to undertake conditional self-binding: it is of utmost importance that standard relations, in which citizens live together as free and equal in freely fortified affirmation of one another's moral status, be restored as swiftly as possible. The asymmetry is thus tenable. We ought not to understand our institutions as enforcing coercively our duties of justice, since such an understanding would de-value our social relationships. Enforcing fortificational punishments if necessary, however, serves to *restore* those relationships.

I am well aware that those who hold a “bottom line” approach to justice—according to which all that matters, from justice's perspective, is that the right distributions of non-interference and resources demanded be secured—are unlikely to be persuaded by all this talk of attitudes. Luck egalitarians, for example, are unlikely to be moved by a concern for the moral quality of our social relations except insofar as they reflect unchosen “psychic costs” for which people must be compensated. The same is true, I suspect, for many libertarians who conceive of all of normative space in terms of property entitlements over one's body and resources. Of greater concern to me here are those who would insist that the state's announcement of coercive threats need not be construed as generative of the sort of debilitating effects on the quality of our relationships that I have suggested. Therefore, I want to end this section by turning to this objection: that I am manifestly mistaken in supposing that a system of deterrent threats would have pernicious effects on our relationships as fellow citizens. Consider the following passage from *A Theory of Justice*:

It is reasonable to assume that even in a well-ordered society the coercive powers of government are to some degree necessary for the stability of social cooperation. For although men know that they share a common sense of justice and that each wants to adhere to the existing arrangements, they may nevertheless lack full confidence in one another. They may suspect that some are not doing their part, and so they may be tempted not to do theirs. The general awareness of these temptations may eventually cause the scheme to break down. The suspicion that others are not honoring their duties and obligations is increased by the fact that, in the absence of the authoritative interpretation and enforcement of the rules, it is particularly easy to find excuses to break them. Thus even under reasonably ideal conditions, it is hard to imagine, for example, a successful income tax scheme on a voluntary basis. Such an arrangement is unstable. The role of an authorized public interpretation of rules supported by collective sanctions is precisely to overcome this instability. By enforcing a public system of penalties government removes the grounds for thinking that others are not complying with the rules. For this reason alone, a coercive sovereign is presumably always necessary, even though in a well-ordered society sanctions are not severe and may never need to be imposed. Rather, the existence of effective penal machinery serves as men's security to one another. This proposition and the reasoning behind it we may think as Hobbes's thesis.⁷⁸

Note that the passage does not reflect any conviction on Rawls's part that coercion is required because people will lack an effective sense of justice and thus will need to be forced to comply with the law; such an admission would compromise the very justification of his theory, given the justificatory significance he assigns to freely willed stability. Note also that Rawls runs together two crucially distinct issues by claiming that "the absence of the authoritative interpretation and enforcement of the rules" is likely to lead to instability. He is certainly right that authoritative *interpretation* is important. John Finnis rightly notes that "the greater the intelligence and skill of a group's members, and the greater their commitment and dedication to common purposes and common good, the *more* authority and regulation may be required, to enable that group to achieve its common purpose, common good."⁷⁹ But as Leslie Green notes: "Coercion secures authority and makes it efficacious, but it does not constitute it."⁸⁰ Authoritative

⁷⁸ Rawls, *Theory*, 211. Emphasis added.

⁷⁹ Finnis, *Natural Law and Natural Rights* (New York: Oxford University Press, 1980), p. 231.

⁸⁰ Green, *The Authority of the State*, p. 75.

interpretation and authoritative enforcement are different. So why does he emphasise enforcement?

The central concern is one of assurance. Now I examined the issue of assurance in Chapter Two. I argued that assurance matters only to the extent that citizens can reasonably demand evidence that their performance of their legal duties will not be *futile*. Now it is not clear from the passage from which the quoted statement is drawn whether Rawls thinks that penal sanctions are a response to citizens' *justified* demands for assurance, or whether they are a response to *pernicious psychological inclinations*, such as envy of free-riders. The latter is certainly implied when Rawls writes: "They may suspect that some are not doing their part, and so they may be tempted not to do theirs."⁸¹ But if the problem is really *informational*—if all people need for the pernicious temptations to lose their practical effectiveness is the *knowledge* that people are complying—the problem can be solved by the "public compliance records" I proposed in Chapter Two, regularly released reports on the quantities and even the names of citizens who are complying and who are not complying.

Public compliance records are a terrific example of motivational fortification: shoring up people's confidence by blunting the practical effectiveness of pernicious temptations. They are also normatively superior to the coercive approach. Some may nevertheless object that there is nothing unobjectionable about the proposal Rawls outlines. Call this *the objection from Democratic Coercion*. Democratic Coercion, let us imagine, is a rival theory to the fortificationist theory of punishment according to which all of us, recognising our envy of free-riders, come together to set up and enforce threats of punishment, simply for the reason of blunting that particular inclination's practical effectiveness. What could be so bad about that?

⁸¹ Rawls, *A Theory of Justice*, p. 212.

My reply is that the mere presence of coercive threats within our fortificational infrastructure impairs our relationships by depriving ourselves of the knowledge that people are choosing to do what is right for the right reasons. To see why, consider Sally. Sally is a fellow citizen in your well-ordered liberal democracy. One day, she runs up to you aggressively and places a gun to your head. You recoil, in fear. She barks: “Don’t even *think* about violating my rights. If you violate my rights, I will harm you. My threats are credible; do not *dare* disbelieve me.” My simple proposition is that this is an extremely morally unappealing way for fellow citizens to relate to one another. And that is true *even* if she adds: “By the way, I’m only doing this to shore up *my* support for just institutions, which is precarious unless I have confidence you will not free-ride...” I would regard my civic relationship with Sally as impaired, because even with her caveat, she has not treated me as someone who has the first moral power: who can see moral reasons and responsibly fortify myself so as to act from them. That I put *my* gun to Sally’s head in response, and utter the same statement, may level the playing field, but it also lowers that field’s absolute moral quality. Hegel’s remark still stings: “To justify punishment in this way is like raising one’s stick at a dog; it means treating a human being like a dog instead of respecting his honour and freedom.”⁸²

One might object that the moral quality of our relations when the state issues anonymous threats is different from the kind of aggressive, interpersonal threatening undertaken by Sally. As B. Sharon Byrd writes: “In threatening punishment, the addressee of the threat is general. An individual need feel ‘restricted’ only if he is considering violating the law, and even then until he violates it the threat remains

⁸² G. W. F. Hegel, *The Philosophy of Right*, translated by T. Knox. (Oxford: Oxford University Press, 1942 [1821]), p. 246.

anonymous.”⁸³ But “anonymous” is misleading. Its literal meaning—“nameless”—is correct: the state does not literally name specific individuals. But *it might as well* name *every* individual. Suppose Sally ran up to *every* citizen, articulating the same threat. Once she has done this to every last citizen, should we then conclude that she has acted anonymously? Simply because it might in some cases be easier to *ignore* the state—say, because fresh technology has rendered the *DRONES* invisible—does not alter the fact that a social order sustained by standing threats of this character is a morally inferior social order. The vision of a government that continually has guns pointed at its citizens’ heads, threatening them with setbacks to their interests if they choose to contravene the norms of justice, defies our understanding of what fully just social relationships involve. What Rawls once called “the bonds of civic friendship”⁸⁴ cannot be realised at the barrel of a gun. If they are, they are no longer those same, intrinsically valuable bonds.

Three caveats are in order. Firstly, this does not mean that general deterrence could not be an effective penal aim in *other* contexts. A system of punishment oriented around the ideal of threatening suffering, and then imposing it, to deter people from acting, could be an appropriate penal theory for a non-ideal *modus vivendi*, or an international system that threatened punishment to rogue states. I do not consider these issues here. Secondly, nothing I have said implies that the police in a well-ordered society should not be *empowered* to enforce laws, for example, if they see someone about to kill somebody and are positioned to stop it. But it matters how we understand the system of punishment; occasional one-off instances of coercive rescue do not imperil the moral quality of our civic relations in the same way standing deterrent threats do. Finally, I am agnostic on the justifiability of *minor fines* attached to particular courses of action as a way

⁸³ Byrd, “Kant’s Theory of Punishment,” p. 184.

⁸⁴ Rawls, *A Theory of Justice*, p. 5.

of reminding or moderately incentivising the agent to choose to do what she knows is right. It is not clear that the social quality of our relations is impacted by the democratic threat of parking fines. I do not, however, take a position on that problem here.

Fourthly, a comment on general deterrence. Crime prevention is achieved by the fortificationist theory of punishment because of the way in which it enables criminal offenders to improve their exercise of their first moral power. This aim serves as its central justification. Many are bound to notice, however, that knowledge about the existence of this institution nevertheless is likely to achieve certain *general* deterrent effects on human beings as we know them. Punishments are, after all, burdensome. Even for less serious crimes that nevertheless merit fortificational punishments, like neighbourhood vandalism, the fact that someone will be required, *inter alia*, to attend Friday evening classes enjoining discussion about the importance of local community, is a prudential reason not to engage in the crime. That citizens have *bound* themselves to the prospect of crime in the manner I earlier described serves only to *increase* the likelihood that people may be at least *somewhat* moved by deterrent considerations: their assigned penal burdens will not be optional civilities, with only the prospect of community disapproval accompanying one's shrugging-off of them. I do not deny this. The sociological function of punishment as a system of general deterrence may well be retained by a system of fortificational punishments. This, I believe, is a cause for regret. But if we manage our public culture in the right way, there is hope that we will be able to understand our system of punishment as centrally fortificational in character. The existence of punishment might thereby serve to deter in exactly the way it should: by reminding us why the criminalised conduct is properly criminalised in the first place.

§6. Disrespecting Agency?

I have addressed several objections throughout the course of the argument. Because the philosophy of criminal law is a subfield in its own right, it is inevitable that in a single chapter I would not be able to cover all of the issues that a full presentation of the fortificationist theory would aspire to address. At least two important issues stand out as beyond my scope despite their importance: firstly, questions about criminalisation—determining the subset of moral duties in political liberalism whose contravention properly triggers a legal duty to submit to fortificational punishment; and secondly, questions about the proper response to offenders who refuse to engage actively with their assigned fortificational exercises, sitting passively. I must save these for another day.

One objection that I must address, however, is the question of whether the fortificationist theory is *disrespectful to the moral agency of offenders*. Does the theory treat them as helpless children rather than responsible adults? There are two central versions of this charge. The first version of the charge condemns the fortificationist theory of punishment as *paternalistic* and, for that reason, clearly wrong by political liberal standards. The second version is more complex, insisting that it is incompatible to claim that a criminal action is both the product of a responsible agent's choice *and* a product of causal forces. I shall take up each in turn.

Deirdre Golash takes to task the broad class of penal theories she terms “moral reform theories,” including R. A. Duff’s proposal that punishment be construed as an effort by the state to engage the criminal wrongdoer in communication oriented to inducing penance and apology, and Jean Hampton’s proposal that punishment’s role is to teach criminal offenders the moral reasons why their conduct is wrongful. Insofar as my view shares a certain restorative impulse with Duff’s and Hampton’s views, it is fair to guess that Golash would advance similar objections against the fortificational theory. She

argues that punishments aiming to reform offenders are wrongful because they are paternalistic: they aim for the good of the offender.⁸⁵ I believe Golash is right to condemn paternalism, and thus her argument would accordingly sting for my view—if *only* the fortificationist theory were paternalistic. It is not, however, paternalistic. To see this, consider a definition of paternalism offered by Quong:

1. Agent A attempts to improve the welfare, good, happiness, needs, interests, or values of Agent B with regard to a particular decision or situation that B faces.
2. A's act is motivated by a *negative judgement* about B's ability (assuming B has the relevant information) to make the right decision or manage the particular situation in a way that will effectively advance B's welfare, good, happiness, needs, interests, or values.⁸⁶

Let us take each prong in turn. I shall read “attempts” to mean “aims to.” Firstly, the fortificationist theory does not *aim* to improve the offender's bundle of goods, or enable her to realise some dimension of her conception of the good, or to increase her subjective state of pleasure. Rather, it demands that she live up to a duty that I have argued she has. But citizens do *not* have any duties to their fellow citizens to fortify their second moral power—the power related to citizens' good. This is principally because free and equal citizens *reasonably disagree* about the nature and proper purposes of their second moral power. Providing one another with equal supporting conditions in which we can then leave one another alone to pursue our particular projects, relationships, and aspirations—whatever they might be—is the essence of the liberal project. Coercing people to satisfy a putative moral duty to fortify their second moral power is absurd because there *is* no such moral duty. But there is, if I am right, a moral duty to fortify one's sense of justice. This marks one crucial disanalogy between the first and second moral powers: we *owe* it to one another to keep our first moral power in good working

⁸⁵ Deirdre Golash, *The Case Against Punishment*, pp. 123ff; Morris, “Persons and Punishment,” p. 122.

⁸⁶ Quong, *Liberalism Without Perfection*, p. 80.

order, whereas we do *not* owe it to one another to keep our second moral power in good working order—indeed, as atheists and devout religious adherents will jointly attest, citizens have no prospects for consensus on what “good working order” might even mean in the second case, and it defies their status as free and equal to expect them to find such consensus.

This leads to the second element of Quong’s definition: the idea that the state should not make *negative judgements* about citizens’ moral powers. But there is a crucial ambiguity between the claim that the state should not make negative judgements about people’s *abilities* (Quong’s formulation) and the claim that the state should not *appraise* citizens’ degree of *effectiveness* in exercising their abilities. Both strike us, properly, as repugnant in the case of the second moral power. This is disanalogous from the first moral power. While it is certainly true that the state disrespects its citizens by claiming that some people’s *capacities* for justice are inferior, it does not disrespect them by claiming that their capacities have been *misused* or *poorly exercised*. I have treaded close to this line at times by using phrases like “defective,” but these terms must always be understood as referring to the agent’s choices about how to *employ* her first moral power—including choices to develop or not develop their skills at exercising it—not the power itself. Moreover, to say that an agent acted wrongly by murdering someone *entails* the proposition they failed to deploy their first moral power adequately.

Thus I conclude that the theory is not paternalistic. But there is a second version of the claim that the fortificationist theory disrespects offenders’ agency. The second version of the objection identifies an important similarity between the fortificationist theory and rehabilitationist approaches to punishment: namely, the interest in *diagnosing* the exact nature of the criminal behaviour with an eye for preventing the likelihood of its

recurrence.⁸⁷ This approach has been widely condemned as incompatible with an aspiration to take seriously the status of criminal offenders as moral agents who were not pushed around by causal forces; they *chose* to act wrongly. Herbert Morris has famously criticised what he terms “the therapy model” of punishment, which he thinks regards instances of wrongdoing as “a symptom of some pathological condition in the way a running nose is a symptom of a cold.”⁸⁸ Morris continues: “Therapy is normally associated with compassion for what one undergoes, not resentment for what one has illegitimately done.”⁸⁹ Jean Hampton, who defends the idea that punishment should teach right and wrong, similarly condemns subjecting offenders to therapy, invoking as a paradigmatic example the practice of electrocuting sex offenders as they view photos of naked children.⁹⁰ This is what happens, she intimates, when punishment is understood to be a form of treatment. “If the psychiatric ‘experts’ decide that powerful drugs, shock treatments, lobotomies or other similar medical procedures are legitimate and necessary treatments of certain criminals,” she asks, imagining the dystopic road to come, “why shouldn't they be used?”⁹¹ In such a world, Golash writes, a criminal offender becomes “a dog to be trained,” rather than an “autonomous human being,”⁹² or, as Morris adds, one of many “animals who must be conditioned.”⁹³

The *modus operandi* of the fortificationist theory of punishment, however, is not one of manipulating the biological or psychochemical constitution of criminal offenders. We have strong reasons to reject such proposals as part of our ideal of what punishment

⁸⁷ Gwen Robinson, “Late-modern Rehabilitation: The Evolution of a Penal Strategy,” *Punishment and Society* 10, 4 (2008): 429-445; B. F. Skinner, *Science and Human Behavior* (New York: Macmillan, 1953), pp. 182ff.

⁸⁸ Morris, “Persons and Punishment,” p. 480.

⁸⁹ *Ibid.*, p. 483.

⁹⁰ Hampton, “The Moral Education Theory of Punishment,” p. 222n.

⁹¹ *Ibid.*, p. 222.

⁹² Golash, *The Case Against Punishment*, p. 127.

⁹³ Morris, “Persons and Punishment,” p. 487.

ought to involve in a well-ordered liberal democracy. Firstly, we have powerful morally-directed pragmatic considerations against assigning states the power to wield such technology, if they had it, on their own citizens. Secondly, the proposition that it is utterly unobjectionable to manipulate, surgically or medically, human beings' biological makeup is likely to be resisted by numerous comprehensive doctrines that we would otherwise classify as reasonable. This alone tells us nothing; it may be that we have duties to undergo medical manipulation that we had not realised, making a doctrine unreasonable if it denies it. But I am not confident we can justify a policy of this kind in a way that does not depend on explicitly or implicitly denying many claims about the nature of human identity and the good that citizens, as free and equal, ought to be able to explore for themselves. At most, such surgical or medical therapy would be optional for offenders who found it congenial to their convictions. But thirdly, and most importantly, *such treatments would not qualify as fortificational in character*. When the character Alex in Anthony Burgess's *A Clockwork Orange* is conditioned so that he experiences incapacitating nausea every time he considers acting violently against others, his sense of justice has not been fortified, but bypassed.⁹⁴ The problem with the conditioning Alex receives is not that it deprives him of his freedom of what to think—he still retains the reflective capacity to affirm convictions⁹⁵—but that it fails to attend to the actual root of the problem: Alex's attitudes toward his fellow human beings. The aspiration of fortificationist punishment, in Alex's case, would be to bring him to reflect on his crimes and develop strategies to do better in the future. But the treatment offered in Burgess's dystopia totally avoids the central problem. And in so doing, it compromises the possibility that Alex ever could have restored relations with his fellow citizens again:

⁹⁴ Anthony Burgess, *A Clockwork Orange* (Harmondsworth: Penguin Books, 1972).

⁹⁵ See the instructive discussion in Rodger Beehler, "Containing Violence," *Ethics* 92, 4 (1982): 647-60.

because his fellow citizens could never have confidence that Alex was honouring his moral responsibilities for the right reasons. Their relationships would be fated to remain impaired.

Nevertheless the objection may still be pressed against the fortificationist theory of punishment that it implicitly denies one of political liberalism's core premises by suggesting, as it seems to do, that criminal offenders "need" additional fortificational measures. How could such a claim be compatible with the notion that criminal offenders are moral agents capable of acting justly on the basis of moral reasons? The apparent tension is assuaged by recalling the simple but all-important distinction between an agent's *likelihood* of acting in accordance with a particular moral standard, and their *capability* of acting in accordance with a particular moral standard. The latter is inescapably guaranteed by our self-understanding as agents; a murderer whose victim's parents respond—"It's all right; you could not have done otherwise"—should feel disrespected; she should *demand* blame.⁹⁶ The former, however, is a matter of our degree of susceptibility to certain influences and pressures; that human beings can exhibit variable degrees of susceptibility of this kind is, as I have said, not remotely controversial.

⁹⁶ I have avoided talk of blame, focusing exclusively on the deliberative function of principles—to guide agents' actions (ought I submit to punishment?)—not the evaluative function of principles. It may seem here, however, that I am relying on an account of blame according to which agents are properly blamed for their actions only if they could have done otherwise, thereby implicitly denying the truth of compatibilist views—a bad move for a political liberal theory aspiring to be ecumenical. I deny this: I am only relying on an account of blame according to which agents are properly blamed for their actions *only if it is an inescapable feature of their proper normative self-understanding that they should take themselves to have been capable of doing otherwise*. That specification is wholly consistent with causal determinism, since it is indexed not to some third-personal account of the causal universe, but to our first-personal, deliberative experience. It allows my view to be robust across compatibilist and libertarian-incompatibilist accounts of blame; indeed, any account of blame deemed "reasonable" by political liberalism must rest in concord with our proper moral self-understanding, the political liberal conception of moral personality. And as I have argued, understanding ourselves to be capable of justice is part of that self-understanding.

The dissolution of this apparent tension, however, has seemed to unearth an even deeper tension. This is the tension between the view of ourselves as reasoning beings, who decide what we do, and the view of ourselves as things kicked about by pressure and temptation. Immanuel Kant understood this to be the defining tension of the human experience.⁹⁷ Thomas Nagel has gone as far as to suggest that this tension—between seeing the world through the first-personal perspective, and the third-personal perspective—is the most fundamental tension in all of philosophy.⁹⁸ Most accounts of punishment have chosen to ignore this tension, embracing one side or the other. Purely therapeutic or medicalised approaches to punishment see criminal offenders as blameless patients who simply need to be treated. Deontological approaches recoil from this picture and run to the other side of the spectrum: if we are to see ourselves as fully free beings, then we cannot let an understanding of ourselves as embodied and socially constituted infect our moral reasoning about what we owe to each other. A retributive theory oriented around the austere moral aspiration simply to *condemn* naturally follows, and it is unrivalled in its philosophical elegance. But from the first-personal perspective of a wrongdoer all-too-aware of her own fallibility and concerned that she take proactive measures to mitigate its influence, condemnation is insufficient. For, we should not aspire merely to *condemn* our fallibility; we should aspire to *do* something about it. I do not thereby *dishonour* my status as an agent by conceding that I am moved by psychological pressure and temptations. We do not disrespect people by predicting their behaviour; if we did, all of social science would be disrespectful. We know full well that we can *both*

⁹⁷ Kant, “The Critique of Practical Reason,” in *Practical Philosophy*, pp. 216ff. I am indebted to the instructive discussion in Christine Korsgaard, *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press, 1996), pp. 200-205, which has helped me to grasp the sheer complexity—but also, I think, phenomenological accuracy—of Kant’s view.

⁹⁸ Thomas Nagel, *The View from Nowhere* (New York: Oxford University Press, 1986), especially pp. 110-37.

offer an account of the social forces that moved people to support Nazism in 1930s Germany while simultaneously endorsing the claim that many such people's actions were the result of their choices—and indeed, that they were morally responsible for choosing otherwise.

David Gauthier said his moral theory was a “moral theory for adults.”⁹⁹ The fortificationist theory of punishment is a penal theory for adults. It does something about fallibility. It takes the striking insight at the heart of both therapeutic approaches to punishment, and Kantian ethics—that wrongdoing is the result of some interference in the successful operation of our reason—and, in a fashion nevertheless congenial to deontologists, saddles individual agents *as agents* with duties to use this knowledge constructively to reduce their susceptibility to such interference. In so doing, the fortificationist theory does not avoid the defining tension in our agential experience. It reckons with it.

Conclusion

Many speak of the purpose of punishment as providing a vital form of security. People have reason to feel safe because they know that those who stand to harm them are afraid to do so. But the achievement of security—the achievement, that is, of a social world in which individuals enjoy a low probability at any given time that they will face a rights-violation—need not be purchased only through coercion and the fear it engenders. By conceiving of punishment as but one piece within a larger fortificational infrastructure,

⁹⁹ David Gauthier, “Moral Artifice,” *Canadian Journal of Philosophy* 18:2 (1988): 385-418, p. 385. Gauthier was referring to his theory's status as a contractarian account that builds morality out of prudential rationality. I am appropriating his slogan for purposes of which, I suspect, he would not approve.

we glimpse the possibility of a form of security achieved not through guns, but through the trust in one another that conscientious fortificational efforts rightly inspire.

CHAPTER SIX

Deliberative Assimilation

“In asking what the most reasonable conception of justice is for a democratic society, we answer a question of considerable importance: we address a disagreement among people who all accept an understanding of persons as equals but who dispute the implications of that understanding. In answering this question, we need not also decide whether the understanding of persons as equals is a compelling cultural assumption, or the most reasonable way to regard people, or a truth of religion or morality.”¹

- Joshua Cohen

“[S]tability means that when tendencies to injustice arise other forces will be called into play that work to preserve the justice of the whole arrangement. Of course, the intolerant sect may be so strong initially or growing so fast that the forces making for stability cannot convert it to liberty. This situation presents a practical dilemma which philosophy alone cannot resolve.”²

- John Rawls

Rawls’s treatment of stability espouses that citizens raised in a well-ordered liberal society will come, firstly, to hold a disposition to act in accordance with standards of justice; secondly; to affirm liberal convictions about what those standards properly are; thirdly, to secure an understanding of their good that will encourage them (or otherwise not interfere with their resolve) to regulate their behaviour by those ideals; and fourthly, to continue to reason effectively about their demands. Only when all four conditions are satisfied for a person, Rawls suggests, may we say that her sense of justice is stable, and only when a sufficient preponderance of a society’s citizens satisfies all four conditions may we say that the justice of the society is morally stable. The arguments advanced in this dissertation have been motivated by the principal worry that even if people acquire a sense of justice educated by liberal ideals, and even if they secure an understanding of their good that coheres with those ideals, they may from time to time nevertheless reach mistaken conclusions about those ideals’ demands that bring them to act wrongly, or

¹ Cohen, “For a Democratic Society,” *Cambridge Companion to Rawls*, pp. 88-89.

² Rawls, *A Theory of Justice*, p. 193.

otherwise face motivationally effective temptations knowingly to advance injustice. Citizens' stability charter, I argued, should have a plan for how to meet these challenges.

Importantly, the standard cases of democratic deliberation, judicial review, and criminal punishment that I have discussed spotlighted the state's treatment of agents who had been socialised within the "moral mainstream" of liberal-democratic society. They may not have had completely reasonable belief-sets about the demands of morality, and they certainly were not motivationally infallible. But I stipulated that they possessed convictions, however practically ineffective or misinterpreted they might have been, that were recognisably liberal: that affirmed ideas like gender equality, liberty of conscience, and the right to free speech. In so stipulating, I was trying to secure a picture of what an ordinary member of a well-ordered liberal society might be like when falling short morally—what you or I might be like, perhaps not at our very worst, but surely not at our moral best. The characters I imagined, like ourselves, were therefore good candidates for fortification. For those suffering motivational failure, such characters were disposed to see their moral weakness *as moral weakness* and resolve to do something about it. And while those suffering from epistemic failure (evidenced by, for example, support for unjust legislation) had, by stipulation, misunderstood the demands of their own convictions, they were in position to come to see the error of their ways—to have their misdirected reason conceivably corrected.

However, liberal-democratic societies contain wrongdoers and prospective wrongdoers that do not fit this familiar description. For some persons within liberal democracies, the values of freedom and equality are not deemed to be properly regulative of all conduct, or indeed any conduct. Call this, for lack of a better term, *severe epistemic*

failure.³ That there are such unreasonable citizens should be clear enough. Without any aspirations of offering an exhaustive taxonomy, we can think of three sorts of examples. Firstly, many seemingly mainstream citizens are committed to the ideals of freedom and equality in broad areas of policy, but decline to grant these ideals their due weight in other areas of policy in which their comprehensive doctrines do not allow such ideals to take precedence over competing values. This accounts for citizens who are happy to adopt decidedly libertarian attitudes in one area of legislation, but insist that their religion disallows them from adopting such an attitude in the area of, for example, marriage law.⁴ Call this *the problem of selective public reasoning*, whereby citizens conscientiously and actively believe that certain cherished principles of freedom or non-interference do not properly apply in a given policy area. Secondly, pockets of religious fundamentalist sects hold unreasonable comprehensive doctrines that explicitly reject the ideals of freedom and equality. The existence of these pockets is made possible by the very liberal protections of free expression, conscience, and association that define the philosophy such groups oppose. Such unreasonable citizens threaten to compromise the rights of people within those communities, as well as imperil the broader polity—both through political organisation that renders the passage of reasonable legislation more difficult, and through terrorist violence. Thirdly, given (a) the plausible assumption that just societies ought to admit some serious number of immigrants, and (b) the difficulty and moral

³ Why is this term less than fully ideal? By describing such agents as having failed epistemically, I am suggesting that their practical reasoning has gone wrong in some way—that they themselves have made a mistaken judgement, deduction, inference, etc. For those who have come to embrace liberal convictions after illiberal pasts, the idea that such individuals were previously mistaken fits their phenomenology. From this it seems to follow straightforwardly that such agents *ought to have known better*. But it is not at all clear that this follows. Often those in deeply unjust contexts are responding intelligently to all the relevant epistemic authorities. Has someone who was caged in a basement for her entire youth “epistemically failed” by not understanding the importance of gender equality?

⁴ Admittedly, it may not always be possible to distinguish certain members of this group from the aforementioned citizens who simply misinterpret liberal ideals, especially when the citizens themselves experience deliberative confusion.

undesirability of “screening” immigrants for reasonableness as they enter our societies, it follows that there is at least a possibility that illiberal immigrants could pose some threat to the enduring maintenance of liberal institutions if they do not achieve the requisite degree of civic assimilation.

How should we respond to the presence of unreasonable agents in our societies? Insofar as unreasonable persons seem particularly likely to engage in such wrongdoing, one proposal for mitigating such threats would be to have the state identify agents with unreasonable convictions and accordingly restrict their rights. The idea, however, that the state should identify and confer subordinate civic status to a certain group on the basis of their convictions is morally unavailable to committed liberals. The alternative response, I contend, must be *communicative*: we ought to engage in a sustained campaign to build and make use of *discursive relationships* with our unreasonable co-citizens, endeavouring to convince them to become more morally reasonable. Rather than restricting or “containing”⁵ the moral agency of our unreasonable brethren, we ought to work to fortify it.

This solution, however, presents political liberalism with an arresting philosophical—and thus practical—challenge. In Chapters One and Two, I explained how political liberalism disavows any official search for its own foundations. A state that is organised under the justificatory auspices of Rawlsian contractualism does not take a stand on *why* seeing and treating others as free equals is morally enjoined, merely that it *is* enjoined. But this generates a serious problem for that state. For, when agents with unreasonable views advance arguments that affront the status of persons as free equals, it is only natural for the political liberal state to counter their denials by *insisting* upon the value of persons as free equals. But when unreasonable agents ask what *grounds* such an

⁵ Rawls, *Political Liberalism*, p. 64n.

insistence, *the state finds itself silent*. It has—seemingly—*nothing to say* to those under its jurisdiction who deny its fundamental values, beyond merely re-describing or re-articulating those values and their demands.

This chapter develops elements of an ethic of deliberation that endeavours to fill that silence, and in so doing help fortify the practical reasoning of unreasonable citizens and, accordingly, bring them into the folds of reasonable consensus. This particular ethic of deliberation—what I call *Deliberative Assimilation*—enjoins citizens to convince their unreasonable co-citizens to become more reasonable. Citizens are subject to duties to look out for unreasonable citizens over whom they could potentially exert a productive influence, and—if they identify them—attempt to convince them through whatever forms of argument are most effective, to abandon their unreasonable convictions. The upshot of my account will be that reasonable adherents to comprehensive doctrines that feature unreasonable sects are likely to be morally required to engage discursively with their unreasonable co-adherents about their political convictions, using *non-public reason* to try to convince them why their shared comprehensive convictions, best understood, demand liberal politics. To save public reason, I shall argue, we must work outside of it.

The structure of the chapter is as follows. In Section 1, I review some discussion on unreasonable citizens in the literature thus far. In Section 2, I turn to the deliberative silence of political liberalism toward its own foundations, and I assess two recent proposals to fill that silence—one by Corey Brettschneider, and another by Micah Schwartzman. In Section 3, I present the fortificational ethic of deliberation I call *Deliberative Assimilation*. At its heart, I shall argue, is *the duty to assimilate*: a duty to attempt to convince unreasonable citizens to become more reasonable by undertaking targeted “discursive rescue missions,” justified by the natural duty to support just institutions. Then, in Section 4, I sketch my proposal that the duty to assimilate may require

reasonable co-adherents of unreasonable comprehensive doctrines to engage in non-public moral reasoning with their unreasonable co-adherents, and I explain why this is both historically and theoretically promising.

Some committed defenders of a free society will doubt the worth of a project of this kind from its very inception. Burton Dreben remarks: “sometimes I am asked, when I go around speaking for Rawls, What do you say to an Adolf Hitler? The answer is [nothing]. You shoot him. You do not try to reason with him. Reason has no bearing on that question.”⁶ But political liberals cheerlead Dreben’s sentiment at their own peril. While it would have been obviously justified to shoot Hitler without first attempting a contractualist conversion—given the dangers he posed—the coercive solution to unreasonableness is insufficient. Seeing citizens as free and equal involves the ascription of the capacity to see moral reasons and act from them. This applies even to Nazis, undermined and misdirected though their moral powers may be. Persuading the unreasonable matters to help secure the (freely willed) justice of our institutional structures. And most importantly, it does so in a manner that engages, rather than bypasses, the moral agency of those to whom we speak. By attempting to fortify the moral powers of others through conscientious engagement and reflection—by trying to stop unreasonableness in its tracks not with guns, but with reason itself—we aspire to prevent injustice in the manner most consistent with respecting those positioned to perpetrate it.

⁶ Burton Dreben, “On Rawls and Political Liberalism,” in *The Cambridge Companion to Rawls*, p. 329.

§1. The Moral and Political Status of Unreasonable Citizens

Who are the unreasonable citizens? They are not simply individuals who happen to make mistakes—epistemic and motivational—from time to time. Were that the relevant criterion, we would all so qualify. Within political liberalism, the term has a much more specific meaning. I shall employ the term “unreasonable citizen” to refer to a member of a liberal-democratic political association⁷ who denies that she and her fellow citizens ought to regulate their social relations by principles that are mutually acceptable among persons understood as free and equal, in light of the significance of the burdens of judgement for that understanding.⁸ Note that this definition emphasises both a normative condition of reasonableness—mutual justifiability among free and equal citizens—and an epistemic condition—recognition of the burdens of judgement.⁹ One common way to misunderstand what an unreasonable citizen *is* involves misunderstanding the content of the epistemic condition. Taking their cue from one common usage of the term reasonableness—“that was a reasonable view of the matter—not irrational or crazy”—as well as some easily misconstrued comments by Rawls—Erin Kelly and Lionel McPherson assume that “unreasonable citizens” are simply the class of

⁷ Immigrants who have not secured legal citizenship are nevertheless properly regarded as members of the political association. For that reason, “unreasonable citizens” can be read to include illiberal immigrants. To be sure, it is reasonable to think that certain rights of full legal citizenship—such as the right to vote—are excluded from immigrants until a certain quantity of time has elapsed. For that reason, it may be that unreasonable immigrants will lack some rights of full liberal-democratic citizenship. But they will lack the rights not because they are *unreasonable* immigrants, but simply because they are immigrants. Rawlsians fresh off the boat are plausibly denied the right to vote as much as anyone else, for reasons I shall not pursue here. For legally-minded normative reflection on immigration and the right to vote, see Gerald M. Rosberg, “Aliens and Equal Protection: Why Not the Right to Vote?” *Michigan Law Review* 75, 5/6 (1977): 1092-1136.

⁸ This explication is consistent with Quong, *Liberalism Without Perfection*, p. 291. As Quong points out, one is properly classified as reasonable even if one rejects only an element of the political liberal ideal, or else refuses to grant it regulative supremacy in her reasoning about what to do.

⁹ Many have interpreted the Rawlsian understanding of reasonableness as divisible along these lines. For example, see Matthew Clayton, *Justice and Legitimacy in Upbringing* (New York: Oxford University Press, 2006), p. 17ff.

citizens whose beliefs are not intellectually respectable.¹⁰ As Jonathan Quong rightly points out,¹¹ this is a gross misreading of the Rawlsian understanding of reasonableness. While reasonableness does include an epistemic element, that element simply concerns recognition of the burdens of judgement—and thus the awareness that others will come to hold reasonable views different from one’s own on such matters as the good when reasoning freely. It does *not* concern whether citizens have sufficient evidence or good arguments to support whatever comprehensive beliefs they avow. Recognising the burdens of judgement, as Quong explains, is not the same thing as endorsing an intellectually respectable comprehensive view.¹²

¹⁰ Erin Kelly and Lionel McPherson, “On Tolerating the Unreasonable,” *The Journal of Political Philosophy* 9, 1 (2001): 38-55, pp. 44ff.

¹¹ Quong, *Liberalism Without Perfection*, p. 295ff. See also Quong, “The Rights of Unreasonable Citizens,” *The Journal of Political Philosophy* 12, 3 (2004): 314-335.

¹² Quong does not pursue the largely biographical question of *why* Kelly and McPherson go wrong in the way they do, but it is worth pursuing for the sake of warning other readers of Rawls against Kelly’s and McPherson’s mistakes. Their confusion, I propose, stems from an apparent asymmetry between Rawls’s account of the reasonable person and Rawls’s account of the reasonable comprehensive doctrine. Each contains moral and epistemic elements. Both the reasonable person and the reasonable doctrine exhibit *moral reasonableness*—commitment to mutual justifiability among free equals. But Rawls’s description of the *epistemic reasonableness* of each is curiously different. When it comes to the reasonable person, as I have said, epistemic reasonableness entails recognition of the burdens of judgement. But when it comes to the epistemic reasonableness of reasonable comprehensive doctrines, Rawls makes some odd comments. He claims that a reasonable comprehensive doctrine employs theoretical reason to describe coherently “major religious, philosophical, and moral aspects of human life” and in so doing “express[es] an intelligible view of the world;” that it employs practical reason, “singling out which values to count as especially significant and how to balance [values] when they conflict;” and that it is “stable over time,” “tends to evolve slowly,” and “belongs to, or draws upon, a tradition of thought and doctrine” (Rawls, *Political Liberalism*, p. 59). This is the most likely basis for Kelly and McPherson’s misinterpretation. In my view, Rawls’s remarks on the character of reasonable comprehensive doctrines are best understood as speculative sociological comments on the sorts of doctrines that actually do persist in societies over time, and thus the sorts of doctrines he believes could join an overlapping consensus on fundamental liberal values. They are not comments on reasonableness as such; if anything, they are comments on *rationality*, and the sorts of doctrines many have used their second moral power to embrace in democratic societies. Rawls is thus oriented here not to reasonable comprehensive doctrines’ *justificatory* role as the sources of citizens’ commitment to freedom and equality, but to their *stabilising* role within an overlapping consensus. But the ambiguity invites the misreading that Kelly and McPherson pursue by suggesting that the idea of reasonable comprehensive doctrines as “intelligible” is somehow equivalent to the idea of recognising the burdens of judgement, which it clearly is not. It is far better to think of reasonable comprehensive doctrines as simply those that comprise

Having clarified one important pitfall in our understanding of unreasonable citizens, we can ask: what ought we to do about them? Unreasonable citizens pose a special problem in existing liberal democracies precisely because they are empowered with all the rights and powers of democratic citizenship. It might be obvious, then, that the best strategy for mitigating their threat to the stability of justice is to *deny* them those very rights. Indeed, on one mistaken but common interpretation of political liberalism, we need not even go to the trouble of *denying* them such rights; for, as unreasonable, they never had any rights to begin with, since (on this reading) they are outside the circle of agents owed justification for the social order. Marilyn Friedman expresses this worry:

If a state's legitimacy depends only on the endorsement of reasonable and rational persons, that means that a state that is endorsed only by its reasonable citizens is thereby entitled to exercise its coercive power over *unreasonable* citizens *without* their consent.¹³

As Quong also explains, Friedman's core mistake—one that Kelly and McPherson also commit—is ascribing to political liberalism the view “that to be worthy of toleration you must be included in the constituency of justification.”¹⁴ But “the constituency of public justification”¹⁵ and the actual constituency of citizens who ought to be respected as free and equal are crucially distinct. In other words, Friedman is eliding the following two questions: Firstly, who are the bearers of moral status? Secondly, who are the idealised reasoners that contractualist theorists imagine when constructing an account of what is involved in respecting the moral status of persons? Note—crucially—that the answer to

reasonable persons, as T. M. Scanlon sensibly emphasises; their epistemic reasonableness is simply their recognition of the burdens of judgement (“Rawls on Justification,” in *Collected Papers*, p. 164).

¹³ Friedman, “John Rawls and the Political Coercion of Unreasonable People,” in *Autonomy, Gender, Politics* (New York: Oxford University Press, 2003), p. 163.

¹⁴ Quong, *Liberalism Without Perfection*, p. 295.

¹⁵ *Ibid.*

the second question is *not* delivered by pointing to a contrasting set of perfectly reasonable citizens who inhabit actual societies and seeing what they happen to endorse. Political liberalism, as I made clear in Chapter One, concerns a *justificatory ideal* that is operationalised fruitfully as a hypothetical contract—a contract that Chapter Three argued we have reason to emulate, for reasons of epistemic fortification—but an *ideal* nonetheless.

If citizens needed to embrace liberal values in order for them to be worthy of liberal values, we would have to conclude that most human beings around the world are not worthy of being treated in accordance with liberal values. Women who are socialised to deny that gender equality is an important moral objective would, in virtue of that socialisation, lack the claim to be respected as equals. This result, however, is pernicious. And it is precisely the perniciousness of this result that has Friedman, Kelly, and McPherson so worried about how we conceive reasonableness. As Quong shows in his analysis, these worries are misplaced once misunderstandings of Rawls's view are set straight. Unreasonable citizens, despite their pernicious convictions, are to be treated as fully equal members of the social order who ought to be free to pursue any conception of the good they so choose within the limits set by justice and entitled to all the liberties and opportunities that render such a pursuit possible. It is therefore false “that the moral force of public reason only extends to those who endorse its premises.”¹⁶

Because unreasonable citizens have the same moral status as reasonable citizens, justice will strongly resist any proposals to circumscribe their rights in response to their unreasonableness. Permanently incarcerating or killing citizens who announce their approval of unreasonable policies is not a morally available option. Many will be inclined to focus on less extreme suggestions—for example, curtailing citizens' rights to speech

¹⁶ Quong, *Liberalism Beyond Perfection*, p. 292. He is disagreeing with the quoted view.

and association with respect to certain injustice-advocating content and activities. I endorse the view, also defended by Quong, that these proposals do not involve rights-infringements at all. Insofar as the interests that underlie the liberties of speech and association—from the perspective of justice—do not include any interest of citizens to advocate and plan for the violation of individuals’ rights, restricting hate speech and the formation of, for example, neo-Nazi political parties, does not actually violate any rights.¹⁷ Indeed, citizens are plausibly under a duty *not* to engage in speech and association oriented to these purposes; without wishing to deny that one can sometimes have a right to do wrong, no one has the right to enact *injustice*.¹⁸ There remain strong moral reasons, however, for our *legally specified rights* not to track our *moral rights* perfectly—for example, because of various costs involved in policing people’s speech and associations. In those cases, fortificational punishments will not attach to those wrongs. I shall assume without argument that the most plausible version of the debate over hate speech laws, as well as laws banning unreasonable political parties, is one about the reasons we have, or don’t have, to make legal rights track moral rights tightly.

Let me therefore turn to a more difficult case: whether unreasonable citizens ought to be stripped of the *right to vote*. Recall my modified endorsement in Chapter One of Eric Beerbohm’s argument that part of democracy’s value flows from the way it makes possible the realisation of just social relationships. But our relationships with unreasonable citizens are not relationships of that kind at all; they are deeply impaired. However concerted our efforts to respect them might be, they are not met with

¹⁷ Jeremy Waldron, *The Harm of Hate Speech* (Cambridge: Harvard University Press, 2012); Quong, *Liberalism Without Perfection*, p. 308ff.

¹⁸ Quong, *Liberalism Without Perfection*, pp. 307ff. Quong’s discussion focuses on Jeremy Waldron, “A Right to Do Wrong,” in *Liberal Rights* (Cambridge: Cambridge University Press, 1993). See also William Galston, “On the Alleged Right to Do Wrong: A Response to Waldron,” *Ethics* 93, 2 (1983): 320-24.

reciprocity. Unreasonable citizens in existing democracies employ the very rights that their civic brethren have secured them to advocate the *elimination* of the very sort of society that has provided them with these rights. How can we insist that they retain political power, when it is clear that they have no intention of using that power in a manner that enables the realisation of the just social relationships whose importance underlay democracy to begin with?

This is a deeper and more difficult question that many democrats appreciate. To use a toy example that helps us clarify the character of our democratic convictions: if I found myself in a situation in which 80% of my fellow citizens were committed to the re-enslavement of some racial minority, and I had the choice of whether to establish democratic procedures or permit the continued rule of a racially egalitarian autocrat, I would face serious moral pressure to choose the latter. The latter clearly involves loss—but the loss, as we saw in Chapter Four, would be accounted for by the horrid circumstances that rendered such a choice appropriate (assuming, without any certainty, that it would be). Democratic decisions to perpetrate injustice are not intrinsically valuable. So how, then, can we insist that unreasonable citizens be afforded democratic rights?

One possibility is to respond in a manner analogous to the discussion of hate speech: all citizens, no matter their convictions, ought in virtue of their basic civic status to be granted a basic right to democratic participation. But such a right entails the power to pursue only *reasonable* laws and policies, not unreasonable laws and policies. Thus while unreasonable citizens have the right to democratic participation, this right does not protect their pursuit of the sorts of political objectives that, as unreasonable, such

persons might be inclined to pursue.¹⁹ Indeed, part of what *judicial review* serves to do is to remind citizens that they are not to use their democratic rights to pursue unjust policies.²⁰ They thus retain their legal voting powers, but are morally (and should be constitutionally) disallowed—as everyone is—from using them to perpetrate injustice. I am satisfied with this response, though I shall not pursue it further here. But even if this response is flawed, there are reasons to doubt that democratic disempowerment of unreasonable citizens would be all-things-considered justified. Firstly, exclusion may well be counterproductive. The assimilationist effects of exposure to public reason could be considerable. By inviting unreasonable citizens to join in the practice of public decision-making, and respecting them as full equals in so doing, we make it more likely that they will reciprocate respect to us. By listening to their grievances, and explaining thoughtfully and courteously why we disagree, we reduce the prospect of alienating and radicalising them—a prospect invited by the proposal to grant second-class citizenship to unreasonable citizens. Secondly, the invasive character of the designation process sets unacceptably risky precedent. While the ideas of reasonableness and unreasonableness are philosophically useful categories for organising our abstract thinking, when turned into rigid legal specifications they become weapons in the hands of self-serving politicians and autocrats. Even in a relatively well-functioning polity, there are bound to be mistakes and abuse in legally designating certain citizens as “unreasonable” and thereby stripping them of their claims to vote. It is far too easy to imagine the matter of designation becoming deeply politicised, with opposing politicians using one another’s

¹⁹ This argument for the right to vote seems to follow from what Quong says, for example, about the fact that we lack rights “to join an explicitly racist political party and endeavour to get it elected.” *Liberalism Without Perfection*, p. 307n.

²⁰ But judicial review is not sufficient to address the problems posed by unreasonable citizens. We must seek to address the problems posed by unreasonable citizens *before* they amass sufficient political clout to enact unjust legislation.

standards of designation against them. Authoritarian regimes would like nothing more than to have access to an immaculately compiled catalogue of their subjects' political preferences. In suggesting that unreasonable citizens could be identified by state police and stripped of their claims to participate in politics, without any risks posed to people's freedom, we arrogantly think we have moved irreversibly from our authoritarian past. Finally, because citizens are seldom wholly unreasonable, restricting them from politics denies them intrinsically valuable opportunities to act morally with respect to the policies about which they *do* have reasonable convictions. Many citizens are committed to the ideals of freedom and equality on a range of policy areas (those I called "selective public reasoners"). And even fully-fledged unreasonable citizens—those who deny the importance of freedom and equality—tend to believe a range of reasonable moral convictions, such as the contention that murdering people for fun ought not to be tolerated. By enabling such persons to advocate for and support such policies, they engage in intrinsically valuable activities of honouring the moral status of others persons.

I have reviewed the state of play in the current literature about whether unreasonable citizens are entitled to full rights, endorsing Quong's position that critics are wrong to accuse political liberalism of stripping unreasonable citizens of their moral rights. Taken together, the balance of reasons rests in favour of the position that all citizens ought to be afforded with rights of democratic participation. Rawls suggests that so long as our society's institutional structure remains essentially just, it will have a natural assimilationist effect on "the intolerant," prompting them to "acquire an allegiance to it over a period of time".²¹ If we learned anything in Chapter Two, however, it is that reasonable citizens are not willing to risk the maintenance of just relations on wishful thinking about human beings' psychological tendencies. Thus, given that

²¹ *Ibid.*, p. 192.

reasonable citizens are properly unwilling to strip unreasonable citizens of their rights, they are moved to pursue some other strategy for mitigating the threat posed by unreasonable citizens to the enduring maintenance of a liberal regime: to engage them discursively and thereby endeavour to convince them to abandon their unreasonable convictions by explaining to them *why* liberal values are worth affirming in the first place. And it is here that we run into a deep and important problem: if political liberalism is silent on the content and nature of its own foundations, how could it persuade anyone to embrace them?

§2. The Silence of Political Liberalism

2.1. The insights and problems of overlapping consensus

To understand and appreciate the silence of political liberalism, we must first understand and appreciate the idea of an overlapping consensus. The overlapping consensus plays two roles in political liberalism, though Rawls is not always clear to distinguish them. Firstly, it reflects the fact that citizens come to endorse fundamental liberal ideals for different moral reasons, each belonging to citizens' respective comprehensive doctrines. The ideal of persons as free equals is the starting point—the first premise—of citizens' shared reasoning.²² Secondly, it helps to stabilise the public order, since each citizen belonging to the overlapping consensus will see justice as consonant with her rational good. And it has further stabilising effects—for example, by giving people assurance that fellow citizens of different religious faiths are serious about justice, and thus increasing people's confidence that diverse strangers across the social order are likely to comply

²² Quong defends the significance of this role of the overlapping consensus in *Liberalism Without Perfection*, p. 180ff.

with justice's demands.²³ (Note that the second role of the overlapping consensus depends on its first role.)

Now the overlapping consensus is a powerful idea for liberal political philosophy. Not only does it reconcile liberalism's affirmation of universal moral principles with its affirmation of widespread moral diversity; it immunises liberalism from the objection that it afflicts citizens with a kind of "moral schizophrenia," forcing them to compartmentalise different beliefs. If a citizen's religion, for example, is itself what instructs her to support liberal justice, then any tension between one's convictions about justice, and one's convictions about the rest of value, disappears. This strategy purchases serious benefits, but at an apparently devastating cost: preventing liberalism from having anything to say to someone who denies the fundamental values of freedom and equality. When the unreasonable citizen asks her legislator, "What moral reason do I have, given my own belief system, to obey your laws?" the legislator can only respond, "Because that is what justice requires—that's what it means to treat people as free and equal." And when she replies, "Why should I treat people as free and equal?" the legislator finds herself without any official, *public* response.

Many have supposed that this is a *virtue* of Rawls's account. Contextualist interpreters of Rawls, such as Richard Rorty, read the overlapping consensus—when juxtaposed with Rawls's invocation that he is reconstructing ideals latent in democratic political culture—as indicating an abdication of liberalism's claim to moral universality.²⁴ This reading, as I mentioned in Chapter One, misunderstands both the rationale for the overlapping consensus and for Rawls's invocations of democratic political culture. As I argued in Chapter Two, a central insight that underlies the overlapping consensus—that

²³ This assurance argument is advanced in relation to the overlapping consensus in Weithman, *Why Political Liberalism*, p. 327ff.

²⁴ Rorty, *Contingency, Irony, and Solidarity* (New York: Cambridge University Press, 1989), p. 57.

it would be *illiberal* for the state to affirm one comprehensive doctrine as the grounds of its central ideals—is itself a claim about the demands of a universal liberal morality. Likewise, the normative self-understanding that Rawls locates as latent in democratic political culture is *our own* normative self-understanding. And it belongs *to* that normative self-understanding that our aspiration to respect others as free equals is an aspiration we have for *everyone*, not simply for those who happen to share that self-conception. Moreover, in an era of unprecedented interaction between cultures—in which immigrants enter our societies and start to demand justification for why the social order is arranged one way rather than another—it is not enough to say, “This is how we do things here.”

2.2. Brettschneider’s “democratic persuasion”

Yet it is precisely this answer that political liberalism seems incapable of providing. How might this silence be filled? What sort of discourse ought to fill it? Focusing on the more specific matter of hate speech, Corey Brettschneider advances the principle of “democratic persuasion” as a middle ground between banning hateful discourse and letting it go utterly unchallenged. But Brettschneider’s account of how exactly the state should discursively challenge hateful speech is inadequate. He argues that the state should endeavour “to convince citizens to adopt the values of equal citizenship.”²⁵ He continues: “[T]he state should non-neutrally express and promote the values of free and equal citizenship. These values form what I call the ‘reasons for rights.’ They explain why we should protect rights in the first place.”²⁶ Thus when the Ku Klux Klan makes the claim that those with black skin ought to be denied, for example, the right to vote, the

²⁵ Corey Brettschneider, *When the State Speaks, What Should It Say?* (Princeton: Princeton University Press, 2012), p. 72.

²⁶ *Ibid.*, p. 73.

state ought to respond by explaining *why* the right to vote is justified: because it is underwritten by the value of free and equal citizenship. “When we seek to explain why such legal protections are afforded to citizens in the legitimate state, we should appeal to the non-neutral, affirmative values of freedom and equality that characterize the status of all who are subject to law.”²⁷

Brettschneider’s strategy to “persuade citizens to adopt basic democratic values” does not address the problem it was designed to solve.²⁸ While it is true that a state that actively *reiterates* the importance of seeing and treating others as free equals could not be seen as *complicit* in the hateful speech it legally allows—a worry that concerns Brettschneider—it is not at all clear that the simple reiteration of such values could *convince haters* to abandon their hateful ways. Those who engage in hateful rhetoric and advocate, for example, that homosexual sex be recriminalised, are well aware that others disagree with them. Illuminating for them why they disagree—because the others buy into an ideal of equality—cannot seriously be entertained as likely to persuade. None of this is to say that the state *ought not* to engage in the sort of speech that Brettschneider advocates. On the contrary, he is absolutely correct when he argues that the state should announce the philosophical justification for its laws, not merely those laws’ content. And we can hope that such announcements’ publicity will have the effect of bending people’s convictions toward the reasonable, prompting “reflective revision” as citizens thoughtfully wonder why their own convictions are at odds with that of the society.²⁹ But we cannot count on it, especially when it comes to citizens who have *conscientiously* adopted illiberal views and would not regard themselves as confused about what morality

²⁷ *Ibid.*, p. 12.

²⁸ *Ibid.*, p. 105.

²⁹ *Ibid.*, p. 52.

requires.³⁰

Brettschneider discusses the possibility that the state should attempt to “go beneath” the ideal of free and equal citizenship to explain *why* this ideal is worth endorsing in the first place, thereby attacking unreasonable agents’ arguments head on. But he explicitly disavows it:

Democratic persuasion could be more effective if it made use of citizens’ own deep ways of understanding the world, including their most deeply held moral views. *But again I want to stress that democratic persuasion cannot promote the values of free and equal citizenship with regard only to effectiveness.* For the state to respect citizens’ freedom of conscience, it must not endorse a particular comprehensive doctrine.³¹

What is puzzling, however, is that Brettschneider contends that both the state “as well as democratic citizens” ought to engage in democratic persuasion, attempting to convince others to adopt fundamental democratic values.³² But it is not clear that a *citizen* fails to respect others’ freedom of conscience when she tries to explain how, for example, her belief in the infinitely rewarding value of love directs her to honour the status of all persons as free equals. This raises the crucial question, ignored by Brettschneider: might *citizens* undertake the work of persuading others to endorse liberal ideals, given their considerably wider discursive prerogative?

2.3. Schwartzman’s conjectural reasoning

Rawls suggests a possibility he calls “reasoning from conjecture,” according to which “we reason from what we believe, or conjecture, may be other people’s basic doctrines, religious or philosophical, and seek to show them that, despite what they might think,

³⁰ Brettschneider does have a tantalisingly misnamed section—“Why Citizens Should Adopt the Values of Free and Equal Citizenship as Their Own”—but it is frustratingly a list of reasons for why *liberals* should care about converting unreasonable citizens to liberal ideals (*Ibid.*, pp. 37–42).

³¹ *Ibid.*, p. 105. Emphasis added.

³² *Ibid.*, p. 4.

they can still endorse a reasonable political conception of justice.”³³ While Rawls does not elaborate this possibility or offer a defence of it, both tasks are undertaken rigorously by Micah Schwartzman. Schwartzman argues “that such reasoning opens up a potentially significant domain of non-public justifications for promoting the ideal of public reason.”³⁴ He notes several examples of reasoning from conjecture in contemporary political philosophy.³⁵ But we can well imagine much more *targeted* efforts at reasoning from conjecture, in which citizens engage one another’s convictions, directly pointing out contradictions or advancing arguments as to why someone’s own premises commit her to liberal conclusions—even as the speaker denies or doubts the truth of the premises themselves.

There is a serious worry that reasoning from conjecture is morally objectionable. Discussing the related matter of whether it is ethically problematic for religious believers to offer a “secular rationale” for a particular policy position even though, as a matter of motivation, they are utterly unmoved by it—and are actually motivated by a sectarian, religious argument—Robert Audi asks:

Does it not smack of manipulation to give reasons that do not move me, in order to get others to do what I want? I use the reasons as psychological levers to produce belief on a basis that does not carry my own conviction. It is as if I invited you to join me on a journey, but by a route I do not sufficiently trust to take it myself. You are entitled to wonder why, if the route is not good enough, or appealing enough, for me to take, I think it good enough for you.³⁶

Schwartzman carefully considers Audi’s assertion here, applying it to the case of

³³ Rawls, “The Idea of Public Reason Revisited,” p. 462.

³⁴ Micah Schwartzman, “The Ethics of Reasoning from Conjecture,” *The Journal of Moral Philosophy* 9 (2012): 521–44, p. 524.

³⁵ Andrew March, *Islam and Liberal Citizenship: The Search for Overlapping Consensus* (Oxford: Oxford University Press, 2009); Lucas Swaine, “Demanding Deliberation: Political Liberalism and the Inclusion of Islam,” *Journal of Islamic Law and Culture* 11 (2009), pp. 92–110; Joshua Cohen, “Minimalism about Human Rights: The Most We Can Hope For?” *The Journal of Political Philosophy* 12 (2004), pp. 190–213.

³⁶ Audi, “The Separation of Church and State and the Obligations of Citizenship,” *Philosophy & Public Affairs* 18, 3 (1989): 259–96, p. 282.

reasoning from conjecture. He notes that Audi elsewhere concedes that one is not manipulative so long as the speaker is clear that she does not accept the premises of the argument she is advancing (though, importantly, she accepts the argument's *validity*). Audi's crucial contention is not that such reasoning is manipulative, but rather: "There is a certain lack of respect implied in seeking my agreement to a policy by offering reasons by which one is not oneself moved."³⁷

Schwartzman argues that while Audi is correct to believe that *public* reasoning must be sincere—a thesis that Schwartzman defends elsewhere with great plausibility,³⁸ and that I endorse—Audi's argument does not apply to reasoning from conjecture. Schwartzman's reason for this "appeals to the fact of reasonable pluralism." He elaborates:

This fact suggests that people may reasonably disagree about doctrines that address matters of fundamental importance, such as the meaning and value of human life. To recognize this fact is to see that people will often have different reasons for coming to similar moral and political conclusions. Thus, it is appropriate in non-public reasoning to assume that others may be motivated by reasons that are different from our own.³⁹

Given the burdens of judgement, reasonable people understand many comprehensive doctrines to be compatible with the normal operation of human reasoning. A citizen may therefore say to another (though perhaps not so gushingly as I have put the point):

I have come through my experiences to affirm what I affirm, and you have come to affirm what you affirm. I can see why it is that you affirm what you affirm, and I respect it, because I respect you as free, though I do not affirm your convictions myself. But there is one thing you affirm that I cannot respect: your rejection of the ideal of persons as free and equal. For despite the fog of mystery that surrounds the meaning of goodness and the nature of the universe, and the scope for disagreement that thereby emerges, one bright light shines through

³⁷ Audi, "The Separation of Church and State," p. 283.

³⁸ Micah Schwartzman, "The Sincerity of Public Reason," *Journal of Political Philosophy* 19, 4 (2011): 375-98. For further discussion of sincerity, see Gaus, *The Order of Public Reason*, pp. 288ff; Micah Schwartzman, "Judicial Sincerity," *Virginia Law Review* 94, 4 (2008): 987-1027; Quong, *Liberalism Without Perfection*, p. 266ff.

³⁹ Schwartzman, "Reasoning from Conjectures," p. 544.

those mists that I insist you must see: the value of each human life. Don't you see it? Don't you sense the importance of respecting the value of every person as equally valuable and in possession of important interests and inspiring capacities? Here—here is one path I can see in your complex array of commitments for coming to see that value. It is not a path I have ever travelled on, but I am gripped by the conclusion of that path—a conclusion that tells us to honour and support each other's capacities to engage freely with the most meaningful questions in life—and I believe that there is great value in you coming to this conclusion, too, for I believe it is true. Please tell me whether you think I am wrong—as an equal, I have no more authority over you than you over me, and I speak to you in the manner I do because I honour and cherish your value as a person, and only want to know why you don't cherish and honour the value of others similarly.

I have tried to elaborate Schwartzman's answer to Audi and to show how it invokes difficult matters at the very foundations of political liberalism. At first glance, Schwartzman's response is persuasive. But there is a second sort of argument to be advanced against Schwartzman that is not dealt with so easily, an argument not concerned with the *pro tanto* acceptability of reasoning from conjecture, but rather with the *effectiveness* of reasoning from conjecture. This is the argument intimated by Audi when he writes that someone on the receiving end of conjectural reasoning “*feels* neither treated with respect nor secure regarding future relations with the persons in question.”⁴⁰ This raises the worry that even if reasoning from conjecture does not strike us as intrinsically inappropriate or disrespectful, it may be *construed* as such by those it addresses.

Why might it be construed as disrespectful despite that it is *not*, in fact, disrespectful? I prepared the answer to this question when I elaborated upon Schwartzman's initial response to Audi. Remember, conjectural reasoning is not *in fact* disrespectful because of the burdens of judgement. But it would not be surprising if unreasonable citizens did not recognise the burdens of judgement (indeed, for a subset, that is analytically true). That bolsters the worry that those who reject freedom and

⁴⁰ Audi, “The Separation of Church and State,” p. 283.

equality will be utterly unmoved by Schwartzman's reasons for why they ought not to feel disrespected when they are on the receiving end of conjectural reasoning—when their deepest convictions are stretched out on the philosophical operating table by an infidel, an unbeliever, or a heretic, who is going to illuminate cheerily where they have gone logically astray.⁴¹

Schwartzman anticipates an objection along these lines when rebutting some criticisms against reasoning by conjecture proffered by Abdullahi An-Na'im.⁴² One of the arguments Schwartzman addresses is that "criticisms made by outsiders are often perceived, and rejected, as forms of cultural imperialism." He continues:

This may give rise to practical reasons for refraining from conjecture. Engaging in it may be pointless or counterproductive. But that some might perceive conjecture as a form of alien imposition is not, in itself, a principled argument against presenting criticisms based on sources internal to their cultural or religious traditions.⁴³

But given that the *point* of reasoning from conjecture is *practical*, not theoretical—to increase the reasonableness of others' convictions, not to complete a scholarly exercise—it matters whether it is going to be counterproductive. If the predictable effect of reasoning from conjecture in certain cases is to radicalise and aggravate unreasonable citizens, the "practical reasons" to refrain that Schwartzman intimates may render reasoning from conjecture all-things-considered unjustified. How is this not a "principled argument"? Perhaps by this Schwartzman means to observe that such an argument, if

⁴¹ This is, of course, a caricature of reasoning from conjecture *done badly*. On the idea that reasoning of this kind ought to have a far greater dialogical character—a point with which I agree and with which I suspect Schwartzman agrees—see Jürgen Habermas, "Religion in the Public Sphere," *European Journal of Philosophy* 14, 1 (2006): 1-25; Jeremy Waldron, "The Ethics of Engaging with Religious Contributions in Public Deliberation," *New York University School of Law Public Law and Legal Theory Research Paper Series*, Working Paper No. 10-84, 2010. Available at <http://ssrn.com/abstract=1708113>. Accessed 5 May 2011.

⁴² Abdullahi Ahmed An-Na'im, *Toward an Islamic Reformation: Civil Liberties, Human Rights and International Law* (Syracuse: Syracuse University Press, 1990), and *Islam and the Secular State* (Cambridge: Cambridge University Press, 2008).

⁴³ Schwartzman, "Reasoning from Conjecture," p. 539.

true, is only true because it is *unreasonable* of the unreasonable citizens to be offended or aggravated in the way they are predicted to be. But that fact does not make the argument false. It merely means it is lamentable that citizens are unreasonable. But we already knew that! It is lamentable that we would even have to *resort* to reasoning to conjecture to get people not to aspire to oppress others. But that does not block Schwartzman's conclusion that we should—*when* it could be productive.⁴⁴

While the practical project of reasoning from conjecture merits further exploration by philosophers and especially citizens, I am unconvinced that conjectural reasoning—whereby members of one tradition tell members of distinct traditions about how the former have misunderstood the argumentative structure of their own tradition—is bound to be productive, or reasonably predicted to be productive, in a sufficient quantity of cases for it to become a major part of the stability charter. If I am right, we have reason to wonder whether there are *supplemental strategies* for convincing unreasonable citizens to be more reasonable. Having spent this chapter surveying what positions are defective, promising, and absent from the current debate, what I want to do is focus in on the particular matter of what members of the *same* comprehensive religious and ethical communities who hold conflicting moral convictions ought to say to one another, for it is the reasonable citizens within such communities who stand to make a productive difference. But first, we need to know more about the underlying moral requirement for which this political strategy is a crucial upshot.

§3. The Duty to Assimilate

The duty to assimilate does not belong to the epistemically fortifying practice of

⁴⁴ Schwartzman's idea of a "principled argument" strikes me as very close to what G. A. Cohen means when he refers to an argument's status as "comprehensively justified." Cohen, *Rescuing Justice and Equality*, p. 41.

democratic discourse that I theorised in Chapter Three. That practice was one designed for basically reasonable, fallible citizens—concerned that they might arrive at some unreasonable policy conclusion or another and thus committed to structuring their collective reasoning and decision-making in a manner that would increase the reasonableness of their conclusions. The ethic of public deliberation grounded in the democracy/contractualism analogy constitutes the mainstream deliberative ethos of a well-ordered liberal democracy. But there are bound to be dark corners even within a well-ordered liberal democracy that, if unfronted, potentially threaten its status as such. For these dark corners, a different deliberative ethic is required: *Deliberative Assimilation*. The fortificational practice of Deliberative Assimilation is one that enjoins citizens to make targeted, communicative interventions into the lives of unreasonable persons with the aim of trying to convince them to become more reasonable—to mount what I shall call *discursive rescue operations*.

What *kind* of a duty is this? The aim of this section is to identify its structure and character. It is not like other duties that belong to the theory of stability. Those duties, derived from what I called *the self-fortification thesis*, were duties to act in ways that increased one's *own* likelihood of accurately identifying and freely complying with the demands of justice. Sometimes, as in the case of democratic deliberation, the most effective way for people to discharge these duties was to work together. But here, the fortificational arrow is pointing in one direction only. If so-called unreasonable citizens were right-thinking enough to *regard* themselves as subject to a duty to submit to a collective practice that enables all to increase their success in identifying principles that treat others in accordance with the political liberal conception of moral personality, they would not actually be unreasonable citizens; they would simply *already* be a part of the mainstream fortificational practice of democratic deliberation.

Nor is the duty to assimilate the sort of duty that corresponds to a determinate right that a particular person might demand. It is nonsense to think that white supremacists have a right to be persuaded not to be unreasonable. That is not simply because of the plain fact that such a demand would be utterly *mysterious* on their illiberal lips. Rather, it is because they—in the context of well-ordered liberal democracy—are properly *blamed* for their possession, and willingness to act on, unreasonable attitudes. That insight is clear from the following case. Jones is a former white supremacist who headed up his local chapter of the Ku Klux Klan. Now retired, he looks back onto his dark past with utter horror. He now affirms the status of his fellow citizens as free and equal. He has since befriended some of the same people whose fundamental political and moral status he long ago impugned, and has expressed his grave regret to them continually. To be sure, it took him a long time to alter his convictions. But now that his convictions are what they are, he is ashamed at his past self for failing to see their veracity. Indeed, he only came to feel the grip of them because of a dear neighbour, Smith. Smith introduced Jones to some young black children. Jones was dismayed at first, but upon coming to learn about the children—what they did for fun, what they wanted to be when they grew up—he came to see them as so *familiar*. It was the beginning of a long road, facilitated by many nights sitting out on the porch with Smith. Now, decades later, Jones runs into Smith. His proper attitude is not: “You were subject to a duty not to violate my right to be persuaded, and you complied; otherwise you would have wronged me.” His proper response—with all due humility, and shame, and a renewed sense of hope in his own first moral power—should simply be: “Thank you for helping me see the force of the moral reasons I have.”⁴⁵ Note how this phenomenology

⁴⁵ Justified expressions of gratitude need not be responses to supererogation. That would run deeply afoul of the phenomenology of moral experience; even in the cases of duties that *do*

helps explain why it is that the duty to assimilate unreasonable citizens is *universally applicable*—it applies *even to unreasonable citizens*. For when Jones reflects back at his moral failures, those moral failures *include* his failures to participate in the cause of shoring up reasonableness. “I should have been *part* of the effort to shore up the freely willed achievement of just institutions,” he thinks, expressing guilt, “not part of the effort to *undermine* them.”

Jones’s response helps us better to understand the kind of duty I have in mind: it is a duty derived from *the natural duty of justice*, which “requires us to support and to comply with just institutions that exist and apply to us” and “to further just arrangements not yet established, at least when this can be done without too much cost to ourselves.”⁴⁶ Because of the dangers posed to fundamentally just arrangements by unreasonable citizens, all citizens are subject to a duty to do what they can to mitigate justice-threatening attitudes. This does not mean the duty does not belong to the theory of stability as I have defined it, since the theory of stability is concerned with the duties that have the purpose of increasing the successful operation of citizens’ first moral power, which this duty clearly aims to do. It just shows that the theory of stability is large and complex; it serves the theory of justice *proper* in many places.

The natural duty to support just institutions yields numerous *derivative duties* in our political and moral life, especially when we find ourselves in non-ideal conditions. My proposal is that *some* of these deliberative duties are duties to engage discursively with unreasonable citizens with the aim of convincing them to become more reasonable. In some cases this duty may even issue derivative duties to engage with unreasonable citizens *in other societies*, but for the purposes of this discussion I fix my attention on the

correlate to particular persons’ rights, such as the duty of a passer-by to rescue a person from drowning, the ritual of expressing gratitude strikes us as perfectly appropriate.

⁴⁶ Rawls, *A Theory of Justice*, p. 99.

derivative duties to engage with one's fellow citizens, including non-naturalised immigrants. The question that arises is this: what are the conditions under which the natural duty to support just institutions triggers a derivative duty for me actually to engage with particular people in a particular manner?

We can locate some answers to these questions within Rawls's discussion of non-ideal theory, which partly involves what the natural duty of justice demands of us in situations that are less than ideal.⁴⁷ Indeed, Rawls himself invites us to imagine what to do if "an intolerant sect comes to exist within a well-ordered society."⁴⁸ There are three criteria that Rawls outlines for the appropriate "politics and courses of action" we ought to pursue when dealing with deviations from the ideal.⁴⁹ Firstly, such policies or courses of action must be "morally permissible." The general matter of moral permissibility concerning our treatment of unreasonable citizens has already been largely addressed. As Rawls writes, "when the constitution itself is secure, there is no reason to deny freedom to the intolerant."⁵⁰ The specific matter of the moral permissibility of engaging unreasonable citizens *deliberatively* has also partly been addressed, when we considered the demand that reasoning from conjecture be "sincere and not manipulative."⁵¹ What other relevant issues might there be? The remaining concern drops out of Rawls's specification

⁴⁷ It is a misunderstanding of the distinction between ideal and non-ideal theory to think that *if* a principle of non-ideal theory applies in a particular situation, it therefore follows that the society in which that situation is taking place is *not* a well-ordered liberal democracy. If that were so, every time someone initiated an attack on someone else that rendered her liable to defensive harm—triggering a situation in which principles of non-ideal theory apply—the society would, in virtue of that single attack, cease to be well-ordered. Without wishing to deny that the attacker would have seriously impaired relations with their fellow citizens as a result of the attack, the characterisation of a society as well-ordered refers to a macro-property of the whole set of social relations. To be sure, continual instances of wrongdoing, especially wrongdoing that alters the basic terms of interaction between citizens, will make the society deviate from its status as well-ordered. But a society does not cease to be well-ordered simply because principles concerning the response to non-compliance apply.

⁴⁸ Rawls, *A Theory of Justice*, p. 192.

⁴⁹ See the discussion in Simmons, "Ideal and Nonideal Theory," pp. 18ff.

⁵⁰ Rawls, *A Theory of Justice*, p. 192.

⁵¹ Rawls, "The Idea of Public Reason Revisited," in *Political Liberalism*, p. 466.

of the natural duty of justice, in which he writes that we ought to take measures to shore up justice, but that we cannot be expected to do so at “too much cost to ourselves.”⁵² Call this the *no unreasonable costs condition* of the duty to assimilate.

The second and third criteria Rawls outlines are that the courses of action must be “politically possible” and “likely to be effective.”⁵³ The question of political possibility is best understood as one of one of *feasibility*.⁵⁴ I shall not investigate this concept here; given that a course of action that is “likely to be effective” is *a fortiori* “politically possible,” I shall focus my attention on the matter of effectiveness. Call this the *effectiveness condition* of the duty to assimilate.

How might we flesh out the content of these two conditions? I intimated earlier that we can think of the specific courses of action enjoined by the duty to assimilate as “discursive rescue missions” to engage unreasonable citizens and save them from their unreasonable beliefs. The flashy designation is partly misleading since, as I have argued, we do not have any duties *to the unreasonable citizens* to persuade them.⁵⁵ This makes such deliberative interventions crucially disanalogous from familiar rescue missions to, for example, rescue drowning children, in which children have the right for passers-by to save them. Moreover, the duty to rescue a drowning child is not a derivative duty flowing from the duty to support just institutions; it is a natural duty in its own right. Nevertheless, pursuing the partial analogy a bit further yields considerable assistance in fleshing out the content of our two conditions. Much of what I say will be quite obvious; my aim is simply to show how our intuitions about classic rescue cases help guide our

⁵² Rawls, *A Theory of Justice*, p. 99.

⁵³ Rawls, *The Law of Peoples*, p. 89.

⁵⁴ For an analysis of the role of feasibility constraints in non-ideal theory, see Holly Lawford-Smith, “Understanding Political Feasibility,” *The Journal of Political Philosophy* (forthcoming).

⁵⁵ Not to them *in particular*, that is; they do not have a *correlative right* to our intervention. We owe it to them insofar as we owe the natural duty to support just institutions to *all* persons.

thinking, to be pursued momentarily, about the duty to assimilate.

Consider, firstly, the *no unreasonable costs condition*. The large literature of the duty to rescue⁵⁶ has continually debated the question of demandingness. It is widely believed that if a putatively required action will require one to sacrifice one's interests to an unreasonable extent, the action is not actually required. The debate falls on what "unreasonable extent," of course, actually entails. There is a large space of possibility between a rescue that moderately burdens me and a rescue that leaves me dead. Debates rage on where to draw the line. Singer's utilitarianism famously commits him to a hard line, requiring one to downgrade the quality of one's life until just above the quality of the person(s) rescued.⁵⁷ Richard Miller proposes a less demanding alternative, according to which any single rescue or general rescue policy that "would impose a significant risk of worsening one's life" is not required.⁵⁸ A middle ground principle more intuitively congenial to political liberalism is offered by Cécile Fabre, according to which a given rescue is unreasonably demanding, and thus not morally enjoined, if it "would prevent the rescuer from leading a flourishing life, where leading such a life means being minimally autonomous, that is, being able to frame, revise and pursue a meaningful conception of the good."⁵⁹ Note that these debates concern two distinct but related

⁵⁶ For example, see Peter Singer, "Famine, Affluence, and Morality," *Philosophy & Public Affairs* 1, 3 (1972): 229-43; David Schmitz, "Islands in a Sea of Obligation: Limits of the Duty to Rescue," *Law and Philosophy* 19, 6 (2000): 683-705; Eric Mack, "Bad Samaritanism and the Causation of Harm," *Philosophy & Public Affairs* 9, 3 (1980): 230-259; F. M. Kamm, "Does Distance Matter Morally to the Duty to Rescue?" *Law and Philosophy* 19, 6 (2000): 655-681.

⁵⁷ "[I]f it is in our power to prevent something very bad from happening, without thereby sacrificing anything else morally significant, we ought, morally, to do it." Singer, "Famine, Affluence, and Morality," p. 235.

⁵⁸ "One's underlying disposition to respond to neediness as such ought to be sufficiently demanding that giving which would express greater underlying concern would impose a significant risk of worsening one's life, if one fulfilled all further responsibilities; and it need not be any more demanding than this." Richard Miller, "Beneficence, Duty, and Distance," *Philosophy & Public Affairs* 32, 4 (2004): 357-383, p. 359

⁵⁹ More precisely, for a would-be rescuer to be exempt from the moral duty, one of the following two conditions must be satisfied: "(1) having to help would deprive him of the all-purpose means

issues. Firstly, they concern the issue of how demanding a *particular* rescue must be before it is no longer morally required. Secondly, they concern the issue of how *many* such rescues one is morally required to undertake before it becomes appropriate to judge that the agent is no longer required to undertake further rescues. I am neglecting this distinction for now, but I shall return to it later.

Consider now the *effectiveness condition* in relation to the classic duty to rescue. The probability that a given rescue effort will succeed matters hugely for determining whether one has a duty to undertake that effort. If the child is swiftly swept away by the rushing current, then even if one were physically incapable of drowning (since one had gills), it is not clear that one would be required to jump in and pursue the rescue. This shows that the duty to rescue is what we can call an *instrumental duty*: a duty that would not obtain were it not for its role in securing some state of affairs—some objective. Instrumental duties are to be contrasted with *intrinsic duties*, such as the duty to praise God for its own sake. So, for example, the duty to rescue a drowning child has the objective of realising a state of affairs in which the child is no longer drowning. All such instrumental duties are subject to the following existence condition:

The effectiveness condition of instrumental duties. A putative instrumental duty is only in fact a duty if there is a reasonable⁶⁰ likelihood that the action that would be required by the putative duty would achieve its objective if performed.

Notice that the satisfaction of this effectiveness condition is both *context-sensitive* and *agent-sensitive*. It is *context-sensitive* because, no matter how good a swimmer one is, if the

necessary to lead any meaningful conception of the good; (2) having to help would deprive him of the specific means necessary to lead the conception of the good which he finds the most meaningful and the pursuit of which makes his life flourishing.” Cécile Fabre, “Mandatory Rescue Killings,” *The Journal of Political Philosophy* 15, 4 (2007): 363–384, p. 365.

⁶⁰ I shall not explain what counts as “reasonable” here, since it is sensitive to issues beyond mere probability. If (1) a putative duty is unlikely to achieve its aspired result, and (2) the costs of trying anyway are extremely small, and (3) the aspired result is extremely important, then I suspect an agent has a duty to try. I ignore this complication for my purposes here since it would take me too far afield to elaborate rigorously the relationship between these three factors.

current is rushing too fast, the rescue attempt will be bound to fail, and is therefore not required. It is *agent-sensitive* because, no matter how slow-moving the current might be, if the on-looking agent cannot swim, “ought implies can” is not satisfied and she has no determinate duty to jump in, as she will not succeed. (The fact that she cannot swim, of course, also implicates the *no unreasonable costs* condition.)

Thus if an agent finds herself in a position in which she stands a reasonably effective chance of succeeding with a rescue effort in a manner that does not pose unreasonable costs to herself, she has a *prima facie* duty to undertake the rescue. I say “prima facie” because things are somewhat more complex. It may be that *several* people satisfy the *no unreasonable costs* and *effectiveness* thresholds, and that if *all* of them attempt to fulfil their putative duties simultaneously, their efforts will be either counterproductive or (in the case of those who end up not actually participating the rescue) unnecessary. Thus in such cases we need an *allocative* criterion to determine the agent upon whom the duty falls. Consider the case of two passing joggers, running by a pond with a drowning child. The chances of each at completing the rescue successfully are above the probability threshold that renders the attempt sufficiently likely to succeed, and neither would risk considerable costs by undertaking. Upon whom does the duty fall? Our considered judgement, *ceteris paribus*, is that the duty falls on the person with the *greatest* chance of success at solving the problem (e.g., the better swimmer). We can call this the *maximal effectiveness criterion of duty allocation*: if there are many people *prima facie* liable to fulfil the duty to rescue, and only one is actually required to discharge it, the burden should fall (consistent with the “no unreasonable costs” condition above, so as to ensure that the best swimmers are not forever called upon) on those most likely to succeed. Consider the

duty to undertake “rescue killings.”⁶¹ Eric has had some minimal practice with throwing knives, but he is not especially good. Masha is a seasoned sharpshooter. *Ceteris paribus*, the determinate duty falls on Masha. The best woman or man for the job gets the duty.

Finally, it is important to notice that *if* the agent is likely to make things *worse* by perpetrating the action, she should refrain from doing so. That is not because of any particular feature of *the duty to rescue*, but simply the distinct *negative* duty to refrain from engaging in harm. It is worth mentioning since foolhardy attempts at rescuing—in cases in which the duty to rescue does not obtain, because the effectiveness condition is not satisfied—can result in the misguided rescuer violating a *negative* duty not to make things worse. If there is a child on a conveyer belt, likely to lose an arm, and by intervening I risk accelerating the speed of the belt, consequently *killing* the child, I should not intervene.

Having completed this brief detour, we are now in position to specify in greater detail the conditions under which people rightly judge themselves to be under a duty to assimilate—a duty, that is, to undertake a discursive rescue mission. *Firstly*, a deliberative intervention should not pose unreasonable costs. People should not have to spend all of their time trying to deliberate with others in an attempt to make them more reasonable. Rather, it should fit into a conception of the good that they deem to make for a personally flourishing life. What this means, too, is that if (a) one believes that one could not live a flourishing life unless one had a strong relationship with one’s best friend, and if (b) one’s best friend held deeply unreasonable views, then (c) *even if* one’s deliberative engagement would have a serious chance of persuading the friend, if (d) it were *also* true that such engagement would be likely to irritate and marginalise the friend and substantially damage one’s relations with him, then (e) one is not required to engage.

⁶¹ Fabre, “Mandatory Rescue Killings.”

Note that (c) is necessary to this formulation because, *secondly*, one would not be subject to the moral requirement to attempt to persuade if there was not a reasonable chance of success. The effectiveness condition of the duty to undertake deliberative rescue means that the duty does not obtain if it is unlikely to accomplish its aim. Moreover—and here is the analogue to the duty *not* to attempt rescue if one might make things worse—if one stands a reasonable chance of *ossifying* or *radicalising* someone’s unreasonable views, given what she knows about the person she is contemplating speaking to, she ought to refrain. It follows from the natural duty to support just institutions that one ought not to act in ways that *exacerbate* the potential for injustice.

Thirdly, the duty should be interpreted in accordance with a *maximal effectiveness criterion of duty allocation*. Suppose you and your friend, Ryan, both know an unreasonable citizen, named Joe. Joe is someone who lives in your own community; whose traditions you are a part of; and whose reasons for rejecting (or declining to apply) the values of freedom and equality you understand and perhaps even once endorsed. Joe’s convictions, however, are alien to Ryan; his reasons for rejecting the ideals of freedom and equality are wholly mysterious to him. Thus, even if we judge that Ryan stands a reasonable chance to make a difference—he passes the threshold for the effectiveness condition—the duty falls on me, not Ryan. We can generalise this case to imagine a division of deliberative labour organised around the following principle:

The principle of maximal deliberative effectiveness: When agents are deciding which unreasonable citizens to engage, each ought to focus on the set of agents over whom she is likely to exert a productive influence and focus one’s deliberative energies on the agents within the set that are most likely to destabilise the achievement or maintenance of justice if they retain their unreasonable convictions.

Notice an unresolved ambiguity. Suppose there are two sets of unreasonable citizens whom I am confident I could engage productively, but I cannot engage both of them

without running afoul of the *no unreasonable costs* condition. To be sure, I would be acting superogatorily if I pursued both. But what if I was intent to choose? Imagine one group is quite small, another quite large. Imagining an implausible precision, let us suppose that my “persuasion probability” with respect to the small group is 0.9, but it is 0.8 over considerably larger group of people—who, by the way, have expressed far greater intent to destabilise just institutions. This example suggests that it is too fast to assume that people ought to expend their limited deliberative energies on those over whom they are *most* likely to be make a productive influence (hence why “most” does not appear in the above principle). So how do we decide?

The key is to remember that the duty to assimilate is itself justified by the prior natural duty to support just institutions. Now, because I read “just institutions” to refer to a social world in which citizens live together as free and equal by freely regulating their civic relationships according to mutually acceptable terms, convincing *any* unreasonable citizen to be more reasonable is a worthwhile accomplishment. Our social relationship with her is impaired, and that is regrettable. But to say that a whole *society* is enduringly just—to say that people’s civic relationships across the social order are fundamentally just—is to make a statement about the character of the whole scheme: the general character of civic relations within the society and thus the extent to which freedom and opportunity are provided for all. Thus, when facing the sort of example I just sketched, one would need to ask oneself about (a) the persuasion probability for each option (*average* persuasion probability in the case of the group), (b) the destabilising potential of each group if left unreasonable. Of course, it is implausible to suppose that we are in position to calculate such figures. There are no algorithms in the ethics of engaging people’s deepest convictions. Such judgements will be precarious, and in many cases extremely difficult. Questions about whether a particular discursive intervention is too

early, risking the ossification of people's already-evolving convictions, are the most difficult of all. In cases in which it is manifestly unclear whom one would be more likely to make a productive influence—when the case is simply unclear—it is morally reasonable to choose either option. In cases in which one is deeply uncertain as to whether one's engagement will be effective or pointless, there is no duty, but it would nevertheless be meritorious to try. And—importantly—in cases in which one's efforts are bound to *fail*, it would be *wrongful* to try.⁶² These distinctions will not map without complication onto the messy world of real politics. But in citizens' hands, they go a considerable distance to helping them organise their thinking about the complex choices they face.

§4. Convince Thy Neighbour:

Comprehensive Communities as Sites of Liberal Conversion

The idea that citizens should articulate their own comprehensive reasons for endorsing particular policy positions is not a new idea. Rawls himself affirms “positive reasons for introducing comprehensive doctrines into public political discussion.”⁶³ Most famous among them are his proviso, according to which “reasonable comprehensive doctrines...may be introduced in public political discussion at any time, provided that in due course proper political reasons...are given that are sufficient to support whatever the

⁶² This is not to suggest that in gravely unjust societies, citizens have a moral duty not to “witness” their protest as they are marched to the gallows for advocating liberal democracy, despite the fact that doing so would harden the illiberal hearts of their enemies. Affirming one's highest convictions in one's final moments is not something someone could reasonably be required not to do; it runs afoul of the “no unreasonable costs” condition. See the very different discussion of “witnessing” in Rawls, “The Idea of Public Reason Revisited,” in *Political Liberalism*, p. 466n.

⁶³ *Ibid.*, p. 462.

comprehensive doctrines introduced are said to support.”⁶⁴ Paul Weithman has persuasively argued that the proviso offers powerful stabilising support.⁶⁵ But because the proviso is restricted to defending particular “principles and policies”⁶⁶—and therefore potentially useful for convincing what I have called selective public reasoners—it is not designed to persuade citizens who reject the fundamental ideals of freedom of equality themselves. It fails to fill political liberalism’s “silence” on that issue. Rawls mentions reasoning by conjecture as one form of non-public reasoning, though does not suggest that it is morally required. A distinct possibility—one I intimated earlier may be superior to reasoning from conjecture—is clear once we see reasoning from conjecture as one potential upshot of a broader moral requirement to engage with unreasonable citizens. My distinctive claim is that political morality sometimes *requires* members of the same comprehensive tradition to work within the non-public terms of that tradition to try to assimilate one’s unreasonable co-adherents into the folds of liberalism.

While Rawls’s specification of reasonable and unreasonable comprehensive doctrines implies that they are, in fact, distinct doctrines, I am confident this here he is simply setting out archetypes, not empirical claims.⁶⁷ Why might we think that reasonable members of comprehensive communities are in a promising position to convince their unreasonable co-adherents? We can offer some plausible sociological assumptions that, if true, corroborate my proposal. Firstly, many of them know each other. Reasonable and unreasonable co-adherents to the same comprehensive doctrine attend church together, collaborate on community projects, and engage in one another’s lives. The bonds of trust are often in place to facilitate awkward and challenging conversations. Secondly, as co-

⁶⁴ *Ibid.*, p. 453.

⁶⁵ Weithman, *Why Political Liberalism*, pp. 330ff.

⁶⁶ Rawls, “The Idea of Public Reason Revisited,” p. 453.

⁶⁷ See Footnote 12 of this chapter on the perplexing character of Rawls’s views about the definition of reasonable comprehensive doctrines.

adherents to a shared comprehensive tradition, they share a framework of texts, traditions, and rituals that give structure to that tradition and its canons of reasoning. Because of their familiarity with this framework, reasonable adherents are more likely to be regarded as “epistemic peers” with respect to the interpretation of the tradition than outsiders.⁶⁸ This becomes especially compelling once we reflect on the fact that deductive arguments are unlikely to be the most promising method, for example, to convince white supremacists or homophobes to abandon their pernicious beliefs. Research in social psychology persuasively demonstrates a fact that should strike Rawlsians as utterly unsurprising: that alteration of deeply held moral beliefs is the result of changes in one’s *considered judgements*. But such alterations are not induced by syllogisms, but often by powerful emotional experiences, discoveries, encounters, flashes of insight, and—most importantly of all—gradual familiarisation with the other.⁶⁹

Some might dispute that the task of persuading unreasonable citizens is best understood as a task *internal* to a comprehensive tradition. Scholars of public reason sometimes invite the picture of a citizen who is deliberating between whether to grant regulative primacy to “political values” as pitted against “comprehensive values.” Rawls himself implies this way of understanding citizens’ beliefs at times. Schwartzman imagines a citizen torn between endorsing the public reasons on a given topic, and the comprehensive reasons, and asking herself: “why should [I] give priority to the values of public reason?”⁷⁰ But this obscures, rather than illuminates, what is going on for such a

⁶⁸ Admittedly, in certain cases, sects that are especially close to one another in interpretation—such as Sunnis and Shi’ites—are *even more* antagonistic toward one another in virtue of how close they are. But it is precisely such cases that make Deliberative Assimilation appropriate, just in a different way: reasonable Shi’ites ought to convince unreasonable Shi’ites to reject their unreasonable beliefs.

⁶⁹ Jonathan Haidt, *The Righteous Mind: Why Good People Are Divided by Politics and Religion* (London: Allen Lane, 2012), especially pp. 27–51.

⁷⁰ Micah Schwartzman, “The Ethics of Reasoning from Conjecture,” p. 523.

person, given the fact of comprehensive heterogeneity, and defies the central insight of the overlapping consensus. When a liberal Christian and an illiberal Christian debate civic marriage equality, political liberalism does not best understand them as arguing between some set of political values and some set of comprehensive values—debating their various weights or determining which comprehensive values to “bracket.” It was precisely this implausible idea of bracketing—which attracted the communitarians’ ire so long ago—that the overlapping consensus was supposed to help political liberalism avoid. It is far more perspicuous to conceive such persons as arguing *among* comprehensive values: namely, whether the best understanding of their comprehensive doctrine vindicates or prohibits a politics of respecting persons as free and equal, including the implications of the burdens of judgement. A comprehensive doctrine is not reasonable simply because it draws enough arrows from a variety of its elements to a variety of liberal-egalitarian public policies. It is reasonable because it affirms the political liberal conception of moral personality and the ideal of a social world in which citizens regard and relate to each other in accordance with that conception.⁷¹

It is important that the argument for this conclusion is predicated entirely on the guidance criteria that I identified and defended in the previous section. It does *not* rest on the view, sometimes defended by political commentators who criticise the reasonable co-adherents of unreasonable persons, that they have special responsibilities simply in virtue of belonging to the same community. But while liberal Muslims may hold a comprehensive conviction that they have special duties to care for other Muslims’ souls, this is not a premise in the argument for Deliberative Assimilation. Rather, just as the morning jogger is in the right place at the right time to be saddled with the responsibility

⁷¹ This is consistent with the rejection of what Quong has called “the common view” of the overlapping consensus, discussed in *Liberalism Without Perfection*, pp. 163ff.

for saving the drowning child's life, the liberal Hindu is in the right place at the right time to convince unreasonable Hindus that proposed legislation in the United Kingdom enumerating a specific provision against "caste discrimination" in existing legislation is morally required. To do otherwise would be to neglect his duty.

The argument that comprehensive communities can serve as sites of liberal conversion is not some arcane proposal from normative political theory. It is a fact of the practice of people's conversion to liberalism. In the American experience with school desegregation, the religious entreaties of Martin Luther King, Jr., were indispensable in securing support for civil rights. In his fictitious "Paul's Letter to American Christians," King speaks in the apostle's voice to his acolytes: "I am impelled to write you concerning the responsibilities laid upon you to live as Christians in the midst of an unChristian world. That is what I had to do. That is what every Christian has to do." He continued:

I understand that there are Christians among you who try to justify segregation on the basis of the Bible. They argue that the Negro is inferior by nature because of Noah's curse upon the children of Ham. Oh my friends, this is blasphemy. This is against everything that the Christian religion stands for. I must say to you as I have said to so many Christians before, that in Christ 'there is neither Jew nor Gentile, there is neither bond nor free, there is neither male nor female, for we are all one in Christ Jesus.' Moreover, I must reiterate the words that I uttered on Mars Hill: 'God that made the world and all things therein . . . hath made of one blood all nations of men for to dwell on all the face of the earth.' So Americans I must urge you to get rid of every aspect of segregation.⁷²

King's religious exhortations increased pressure from Christian communities, black and

⁷² Martin Luther King, Jr., "Paul's Letter to American Christians," *Delivered at Dexter Avenue Baptist Church, Montgomery, Alabama*, 4 November 1956, available at http://mlk-kpp01.stanford.edu/index.php/encyclopedia/documentsentry/doc_pauls_letter_to_american_christians; accessed 14 January 2013. King also advocated for distributive justice in these speeches using religious speech: "God never intended for one group of people to live in superfluous inordinate wealth, while others live in abject deadening poverty. God intends for all of his children to have the basic necessities of life, and he has left in this universe 'enough and to spare' for that purpose. So I call upon you to bridge the gulf between abject poverty and superfluous wealth."

white alike, to participate in desegregation efforts. Today, liberal churches stand to play an indispensable role in accelerating the movement for LGBT rights.⁷³

Importantly, such communicative engagements are not best understood as necessarily consisting in *syllogisms*. One of the great takeaways of the idea of the overlapping consensus is that there are no clearly established, common standards of reasoning by which everyone either did or could become a liberal. Art and storytelling play paramount roles. Moreover, while I have emphasised the importance of getting people to respect human capacities as the key to embracing liberalism, this is only one argumentative path. Consider that certain peace-cherishing comprehensive doctrines might begin with the thought that a *peaceful* social order is the most desirable one. But it then recognises the burdens of judgement, seeing that the free exercise of practical reason will lead people to different conceptions of the good. Therefore the only way to sustain a political society predicated on a single conception of the good would be through the oppressive apparatus of the state—a recipe for civil strife. Therefore, we might decide that we should permanently embrace the political liberal conception of moral personality, not because of the intrinsic worth of so doing, but because it is the only way to avoid a political culture of violent oppression.

One of the most inspiring developments in contemporary efforts to convince unreasonable citizens to abandon their violent practices has been the willingness of *former* unreasonable citizens—those who have since embraced liberal ideals—to take the lead in

⁷³ Gary Hall, dean of the Washington National Cathedral—the go-to church for U.S. presidential funerals—has recently initiated his support for same-sex marriage, marshalling his pulpit to advance religious arguments for non-discrimination. “I read the Bible as seriously as fundamentalists do,” said Hall. “And my reading of the Bible leads me to want to do this because I think it’s being faithful to the kind of community that Jesus would have us be.” Qtd. in Joe Coscarelli, “National Cathedral to Get In on Gay Weddings,” *New York Magazine*, 9 January 2013, available at <http://nymag.com/daily/intelligencer/2013/01/national-cathedral-to-get-in-on-gay-weddings.html>, accessed 20 January 2013.

engaging their unreasonable brethren. While not strictly speaking members of the same comprehensive doctrines, former gang members have been indispensable players in the movement to convince adolescent males to abandon participation in violent gangs—an illustration of my general thesis.⁷⁴ So, too, have former Islamic extremists played key roles working with young males to convince them to abandon their beliefs on the justifiability of certain violent tactics.⁷⁵ In some cases, defections from Al-Qaeda have been traceable to conversations between prospective suicide bombers and their family members.⁷⁶ And former skinheads and white supremacists are actively involved in the

⁷⁴ One may doubt that those who participate in violent gangs are incurring epistemic failure, as opposed to simply motivational failure. Based on testimony of members, it is relatively clear that while many regard their situations as decidedly non-ideal, they nevertheless resolve that violence against their gang rivals and their rivals' loved ones is justified, given the circumstances. Some may insist that the urban ghetto resembles a Hobbesian state of nature, in which everyone has the right to do anything to survive; but the claim that people have a moral permission to do *anything* to survive is highly implausible. Given the number of innocent non-participants who die from gang violence, the attitude I just described on the part of gang members can only be classified as unreasonable. I do not discuss gangs here, despite the relevance of the topic to the theory of stability, because I am hopeful that *this* brand of unreasonableness—generated by the peculiar, injustice-engendered conditions of the modern urban ghetto—would be significantly less common in a well-ordered liberal democracy with equal educational and economic opportunities for all. For further discussion of gang conversions, see Scott H. Decker, David C. Pyrooz, *Leaving the Gang: Logging Off and Moving On* (Washington, DC: Council on Foreign Relations, 2011).

⁷⁵ Jessica Stern, "Mind Over Martyr: How to Deradicalize Islamist Extremists," *Foreign Affairs* 89, 1 (2010): 95-108; Angel Rabasa, Stacie Pettyjohn, Jeremy Ghez, Christopher Boucek, *Deradicalizing Islamist Extremists* (Arlington: RAND Corporation, 2010); Renee Garfinkel, "From Religious Extremist to Peacemaker," *Common Ground News Service: Constructive Articles that Foster Dialogue*, 19 June 2007, available at <http://www.commongroundnews.org/article.php?sid=1&id=21200>, accessed 5 January 2013; Christopher Boucek, Shazadi Beg, and John Horgan, "Opening Up the Jihadi Debate: Yemen's Committee for Dialogue," in *Leaving Terrorism Behind: Individual and Collective Disengagement*, edited by Tore Bjørgo and John Horgan (New York: Routledge, 2009); Richard Barrett and Laila Bokhari, "Deradicalization and Rehabilitation Programmes Targeting Religious Terrorists and Extremists in the Muslims World: An Overview," in *Leaving Terrorism Behind*; Maajid Nawaz, *Radical: My Journey from Islamic Extremism to a Democratic Awakening* (London: W. H. Allen, 2012); Walid Ramzi, "Algerian Imams Engage Militants in Dialogue," *Magharebia*, 24 August 2012, available at http://magharebia.com/en_GB/articles/awi/features/2012/08/24/feature-02, accessed on 15 January 2013; Naureen Chowdhury Fink, and Hamed El-Said, *Transforming Terrorists: Examining International Efforts to Address Violent Extremism* (New York: International Peace Institute, 2011).

⁷⁶ Rajeh Said, "Why al-Qaeda members defect," *Magharebia*, 24 November 2011, available at http://magharebia.com/en_GB/articles/awi/features/2011/11/24/feature-02, accessed on 15 January 2013; Jason Burke, "Senior extremists linked to al-Qaida, surrenders to Saudi

movement to convince current members to defect—in politics from the U.S. to Sweden.⁷⁷ It is plausible to think that such persons have *especially* strong duties to engage in such work, to *rectify* the wrongs they have perpetrated in the past (though I do not defend that proposition here). The strategies of engagement used by all—enjoining persons to reflect on what they value and why, asking them whether violence will achieve their purposes, and providing them an alternative community of people to support them through their decisions to abandon unreasonableness—are remarkably similar across all three sorts of cases.⁷⁸ Of course, what these cases teach us is an important lesson. Unreasonable citizens are dangerous; they avow their convictions sincerely and may be as keen to convince you as you are keen to convince them. Every deliberative encounter brings risk. Unreasonable citizens are not children to whom one must clarify, in simple terms, how they have gone wrong. They are agents with beliefs, and any self-respecting agent is bound to argue back when their deepest convictions are challenged. That is precisely why the bonds of support that already exist among people with similar life experiences must be harnessed, in precisely the way I am suggesting, to shore up the ideals of freedom and equality in our communities. The attitudes of former unreasonable citizens—many who dedicate themselves to helping others “make the escape” from

authorities,” *The Guardian*, 5 May 2011, available at <http://www.guardian.co.uk/world/2011/may/05/extremist-alqaida-surrenders-saudi-arabia>, accessed 15 January 2013.

⁷⁷ T. J. Leyden and M. Bridget Cook, *Skinhead Confessions: From Hate to Hope* (Springville: Cedar Fort, Inc., 2008); Tore Bjørgo, Jaap Van Donselaar, and Sara Grunenberg, “Exit from Right-Wing Extremist Groups: Lessons from Disengagement Programmes in Norway, Sweden, in Germany,” in *Leaving Terrorism Behind*.

⁷⁸ Rony Berger and Philip Zimbardo, “Creating a Partner: A Qualitative Study of Political Extremists and Ex-Gang Members Who Have Chosen the Anti-violence Path,” Commissioned by Google Ideas, March 2012, available at <http://ccare.stanford.edu/?article=creating-a-partner-a-qualitative-study-of-political-extremists-and-ex-gang-members-who-have-chosen-the-antiviolence-path>; accessed 15 January 2013; Maajid Nawaz and Timothy Zaal, “The Road to Damascus,” *Global: The International Briefing* 9 (2012), available at <http://www.global-briefing.org/2012/01/the-road-to-damascus/>, accessed 10 January 2013; Jared Cohen and Brendan Ballou, *Becoming a Former: Identity, Ideology, and Counterradicalization* (Washington, DC: Council on Foreign Relations, 2013).

unreasonableness—testify to the conviction that all citizens are properly tasked with nourishing the conditions of justice wherever we find them lacking.

Conclusion

The approach in this chapter aims to remedy a prevailing asymmetry within liberal political philosophy. If I am an agent who endorses the values of freedom and equality and I seek guidance on how to relate to and engage fellow agents who share those convictions, there is a considerable amount of political philosophy to assist me. But if I am a citizen who endorses those values who wants to know how to relate to and engage fellow citizens who do *not*, I am out of luck. Political philosophers are nearly silent. Deliberative Assimilation aims to help fill this void. In so doing, it is an example of twenty-first century political philosophy: political philosophy that accounts not simply for how committed liberal-democrats engage with one another, but how they engage reflectively with the world beyond.

CONCLUSION

In this dissertation I have argued that a commitment to justice properly directs citizens to design and submit to measures that increase the likelihood that they will accurately identify and freely comply with the norms of justice that apply to them. I have explained that the Rawlsian ideal of stability properly plays a powerful and important role in our normative political theory and democratic practice—simply not the one it was initially cast by Rawls to play. By conceiving stability as a practical challenge set for us by justice itself, rather than a condition on what justice is, our empirical understanding of our nature and its apparent weaknesses ceases to be a threat to our moral ideals; rather, such knowledge empowers us to pursue those ideals with greater intelligence—and in a manner that better coheres with the nature of moral commitment and our self-understanding as agents. Finally, the view I have developed helps us secure deep and morally compelling justifications for many of our most familiar and controversial practices—from democratic deliberation to judicial review, from criminal punishment to our civic encounter with “the other.”

Throughout the dissertation, I have deliberately taken up the first-person perspective of idealised citizens of a well-ordered liberal democracy: a society in which persons live together as free and equal by regulating their social relations by mutually acceptable terms of interaction. In so doing, I have fixed my gaze only on a handful of examples from just one element of the theory of stability: the question of whether a fundamentally just social order faces risks of moral degeneration. Nevertheless, my general thesis—that citizens ought to take measures to fortify their moral reasoning and motivation—applies to all moral agents, here and now. The ideal of a stability charter for a well-ordered liberal democracy thus serves as an ideal to which *we* should strive in our

presently less than fully just circumstances. Each proposed element of the stability charter I have discussed has novel implications for contemporary political practice.

Firstly, our democratic deliberations ought to be structured to improve the collective success of our moral reasoning. Citizens have moral reason to work harder to be open-minded about the content of their moral commitments, orienting their shared reasoning around the ideal of public justification among free and equal citizens. By conceiving their collective deliberations as attempts to realise the very process of intersubjective justification through which norms of justice are constructed—taking seriously Rawls’s exhortation to think of those norms as “the result of an agreement”¹—citizens can fortify their moral reasoning and simultaneously “reignite the radical democratic embers of the original position in the civic life of their society,”² continually reaffirming one another’s status as free and equal.

Secondly, the practice of rights-based judicial review must be reformed. While the vision of *ex ante* weak foundational judicial review that I have defended is politically infeasible at the federal level in the United States, owing to entrenched legal arrangements, there is no reason my companion account of *ex post* strong foundational judicial review should not become the animating ideal of the U.S. Supreme Court’s approach to rights-based judicial review of legislation, as well as the ideal of any other court that takes up *ex post* rights-based judicial review. Such courts only have reason to strike down legislation when it is beyond the pale of reasonable disagreement, and to do so in a manner that fortificationally engages the democratic citizens whose moral fate such courts seal. For nations writing fresh constitutions, I have offered guidance as to

¹ Rawls, *A Theory of Justice*, p. 48.

² Habermas, *The Inclusion of the Other*, p. 69.

how to pursue constitutional design in a manner that both orients democracy toward the achievement of justice while sustaining the moral quality of just civic relationships.

Thirdly, the existing practice of criminal punishment ought to be reformed to greater resemble the system imagined by the fortificationist theory of punishment. Existing systems of punishment across liberal democratic societies are justified through a combination of desert retributivism, according to which the guilty deserve to suffer, and deterrence-based measures, according to which we subject citizens to coercive threats to alter their prudential calculations. Neither view rests in accordance with our self-understanding as free and equal citizens, capable of justice. By insisting that individual offenders are morally tasked with pursuing appropriate measures to reduce the likelihood that they will offend again, we conceive of punishment as a productive course of action through which offenders reckon with their past to shore up their commitment to the future. Existing practices of restorative engagement between victims and offenders, community service orders, education in prisons, and cognitive behavioural therapy ought to be radically expanded.

Finally, the silence of political liberalism toward its own foundations must be reconceived. While the political-liberal state—the collective organisation that speaks and acts in the name of all—is properly silent on its own foundations, liberal citizens must themselves take up the task of convincing those who reject liberal ideals to reconsider their position. Especially given the prodigious degree of selective public reasoning in existing liberal democracies—often in cases in which religion either motivates the unreasonableness, or could be harnessed to alleviate it—co-adherents to the same comprehensive doctrines ought to take up the task of persuading their co-adherents to see the importance of respecting all as free and equal.

My approach has thus generated important insights that can guide political and moral reform in non-ideal societies. I have only scratched the surface, however, of the research agenda in political philosophy that the theory of stability enjoins us to pursue, especially when it concerns the precarious task of transitioning from unjust societies to just societies. Several further avenues of research promise to deepen and extend our understanding of the theory of stability—helping us to craft a stability charter not for the well-ordered society, but for our own.

Firstly, in an age of burgeoning diversity, immigration, and interaction between liberal and illiberal societies, the silence of political liberalism is no longer problematic simply for our domestic quarrels. We can no longer afford to assume, falsely, that the natural duty of justice generates duties for the maintenance of one's home institutions alone. In a globalised world, Deliberative Assimilation must go global. What deliberative norms properly govern the relations between citizens of liberal societies and citizens of illiberal societies? Do citizens have duties to harness their religious and ancestral connections with other societies to assist in improving the reasonableness of the citizens within those societies? Important questions of "global non-public reason" remain unanswered.

Secondly, while I have attended in this dissertation to the question of the first moral power's successful maintenance, this is only the second half of Rawls's concern with stability. A liberal democracy's stability charter also requires an account of that moral power's cultivation through childhood. Political liberals have tended to do well on this front—offering theories of civic education that suit the citizens of the well-ordered society. But in our non-ideal circumstances, political liberalism desperately requires a theory of civic education for unjust circumstances. How should individual teachers, school administrators, parents, and officials who find themselves striving to cultivate the

sense of justice of schoolchildren in the face of perverse incentives and malfunctioning institutions act? Citizens need normative principles to short-circuit the reproduction of unjust arrangements by educating the citizens of the future out of, and away from, the sort of thinking responsible for such reproduction, and normative political theorists need to provide those principles. How can we teach people to view themselves and others as free and equal when raised in a social context that fails to communicate that message?

Finally, normative political theory needs to address the moral challenges raised by a phenomenon we can call *subversion*. Subversion transpires when citizens or officials act in ways that have the effect of undermining the successful development or exercise of others' moral powers. In the case of contemporary urban ghettos, unjust public policies generate conditions in which citizens face a striking incongruence between what the criminal law requires of them and what they think they need to do in order pursue fundamental human goods of income, security, and belonging. That the ensuing tempest of gang violence is indeed the partial result of political injustice is not controversial. But if I am right that citizens have fortificational duties to work together to sustain a context that supports all in freely complying with the fundamental norms of justice, this state of affairs reflects a further, under-theorised wrong: shutting citizens out of our shared fortificational infrastructure and thereby depriving them of propitious conditions in which to pursue and enjoy just relationships. Normative political theory needs to secure a better understanding of this injustice—both the nature of its wrongness, and its effects (such as cultivating false beliefs about what we owe to each other). It also needs to understand the implications of this injustice for the moral standing of the state to punish offenders. Some defenders of communicative theories of punishment, for example, have argued that a government that subjects its citizens to injustice has forfeited its moral authority to hold them accountable when they act wrongly. Does the fortificationist

theory of punishment face a similar problem? Only by taking the theory of stability into the darkest territories of non-ideal theory can we secure an answer.

These problems are only some of the urgent many that a full theory of stability would address. But no matter what research avenues are opened up by the theory of stability, one abiding conviction unites them all: the belief that it is possible for human beings to do better than we have done when it comes to respecting one another as free and equal. No one forfeits her moral status in virtue of having made a moral mistake, however serious. Even if someone acts wrongly time and time again, part of what it *is* to see her as a moral agent is to see her as capable of reform. We must never give up on others.

I want to close by reflecting on that sentiment—for to many, it will appear to be *nothing more* than sentiment. The notion that the foes of justice are properly treated as outside the ambit of justice's concern is one with a venerable pedigree in political philosophy. Thomas Hobbes argues that “if a subject shall by fact, or word, wittingly, and deliberately deny the authority of the Representative of the common-wealth,” it is only appropriate—as “an enemy of the Commonwealth”—that such a person “lawfully be made to suffer whatever the Representative will.”³ They cannot simply be punished for their transgression, Hobbes believes, for by transgressing in the way they have they have removed themselves from the community of subjects, placing themselves “in declared Hostility”⁴—at *war*—with the state and its people. John Locke develops this theme even more arrestingly, writing:

One may destroy a man who makes war upon him, or has discovered an enmity to his being, for the same Reason, that he may kill a *wolf* or a *lion*; because such men are not under the ties of the common law of reason, have no other rule, but that of force and violence, and so may be treated as beasts of prey, those

³ Hobbes, *Leviathan*, p. 356.

⁴ *Ibid.*

dangerous and noxious creatures, that will be sure to destroy him, whenever he falls into their Power.⁵

There are, to be sure, numerous exegetical questions to be asked about this passage and what, exactly, Locke meant by it.⁶ This may simply be dramatic language for the widely affirmed claim that persons can render themselves liable to incur certain forms of defensive harm in virtue of wronging or posing threats to others. But there is a more insidious impulse that we can recognise in Locke's passage, whether he intended it or not. It is an impulse we see in public debates about criminal punishment, and about the proper ways to deal with threats posed to our way of life from terrorism. This impulse recognises the high-mindedness of just social relationships; it does not *deny* that the world would be wonderful indeed if our public life were marked by widespread trust in one another. But high-minded ideals make for dangerous politics. On this view, the stability charter I have advanced is one destined to disappoint—an artifice of timid and naïve philosophers who talk a big game about human fallibility but, in the end, refuse to shrug their utopianism. But such utopianism, the sceptics say, is suicidal. Those who threaten human lives and just institutions must not be handled gently. They are too dangerous. That we are epistemically fallible only makes such figures—the criminal and the unreasonable among us—an even greater threat. Liberal rules do not apply to those who themselves refuse to play by them. With teeth gritted and fists clenched, we must bury our pride and set about “containing them—like war and disease”⁷—in the most literal reading of that phrase.

⁵ Locke, *Second Treatise of Government*, p. 14.

⁶ See especially John Dunn, *The Political Thought of John Locke: An Historical Account of the Argument of the 'Two Treatises of Government'* (Cambridge: Cambridge University Press, 1969), pp. 107ff; A. John Simmons, *The Lockean Theory of Rights* (Princeton: Princeton University Press, 1994), pp. 148ff; Jeremy Waldron, *God, Locke, and Equality: Christian Foundations of John Locke's Political Thought* (Cambridge: Cambridge University Press, 2002), pp. 141ff.

⁷ Rawls, *Political Liberalism*, p. 64n.

The thesis that non-liberals ought not to be respected in accordance with liberal standards is not one, however, that our moral self-understanding can tolerate. The proposition that any given agent presupposes herself to be capable of justice is only part of the picture. By the sheer act of engaging in moral criticism of others, we imply that they are capable of justice. Our reactive attitudes presuppose that the harm perpetrated by human beings is qualitatively different from the harm done by tornadoes or falling rocks. And the explanation of *why* is that it is simply inescapable to see human beings as creatures who *reason*. To be sure, we do not always reason correctly or live up to reason's demands. But it is precisely because we assume that everyone *could* do so that we resent one another's unreasonableness in a way we would never resent the conduct of a wild animal.⁸ It is precisely because we *believe* in people's capacity to live up to those standards that we are upset with them when they fail to do so. "Blame," writes Christine Korsgaard, "...declares to its object a greater faith than she has in herself."⁹ To be constantly doubting whether people are capable of justice, in every encounter, so that one can determine whether they have a claim to be treated as free and equal, is to live one's life as though continually walking through a world amidst monsters that might turn and destroy you at any moment. But few people would regard such a life as one worth living. This is not to deny that treating others as responsible agents is a risky business:

To hold someone responsible is to regard her as a *person*—that is to say, as a free and equal person, capable of acting both rationally and morally. It is therefore to regard her as someone with whom you can enter the kind of relation that is possible only among free and equal rational people: a relation of reciprocity. When you hold someone responsible, you are prepared to exchange lawless individual activity for reciprocity in some or all of its forms. You are prepared to accept promises, offer confidences, exchange vows, cooperate on a project, enter a social contract, have a conversation, make love, be friends, or get married. You are willing to deal with her on the basis of the expectation that each of you will

⁸ See the discussion of reactive attitudes in P. F. Strawson, "Freedom and Resentment," in *Freedom and Resentment and Other Essays* (Abingdon: Routledge, 2008), pp. 6ff.

⁹ Korsgaard, *Creating the Kingdom of Ends*, p. 196.

act from a certain view of the other: that you each have your reasons which are to be respected, and your ends which are to be valued. Abandoning the state of nature and so relinquishing force and guile, you are ready to share, to trust, and generally speaking to risk your happiness or success on the hope that she will turn out to be human.¹⁰

The risks and rewards of relating to one another as free and equal do not come separately. While the dangers can be mitigated through hard work, they will never be eliminated; human beings will always remain free to choose wrongly, so long as they remain human. The only question to ask ourselves is: what sorts of beings could have such extraordinary value as to make the dangerous journey to justice worthwhile?

¹⁰ *Ibid.*, pp. 189-90.

BIBLIOGRAPHY

- Abizadeh, Arash, "Cooperation, Pervasive Impact, and Coercion: On the Scope (not Site) of Distributive Justice," *Philosophy & Public Affairs* 35, 4 (2007): 318-358.
- Ackerman, Bruce, *Social Justice and the Liberal State*, New Haven: Yale University Press, 1980.
- , *We the People, Volume 1: Foundations*, Cambridge: Harvard University Press, 1991.
- , *We the People, Volume 2: Transformations*, Cambridge: Harvard University Press, 1998.
- Ackerman, Bruce and James Fishkin, *Deliberation Day*, New Haven: Yale University Press, 2005.
- Alexander, Lawrence, "The Doomsday Machine," *The Monist* 63, 2 (1980): 199-227.
- Alighieri, Dante, "The Inferno," in *The Divine Comedy*, translated by Henry Wadsworth Longfellow, Leipzig: Bernhard Tauchnitz, 1867.
- Allison, Henry E., *Kant's Theory of Freedom*, Cambridge: Cambridge University Press, 1990.
- Americks, Karl, "Kant, Hume, and the Problem of Moral Motivation," in *Kant and the Historical Turn: Philosophy as Critical Interpretation*, New York: Oxford University Press, 2006.
- An-Na'im, Abdullahi Ahmed, *Toward an Islamic Reformation: Civil Liberties, Human Rights and International Law*, Syracuse: Syracuse University Press, 1990.
- , *Islam and the Secular State*, Cambridge: Cambridge University Press, 2008.
- Anderson, Elizabeth, "What Is the Point of Equality?," *Ethics* 109, 2 (1999): 287-337.
- Anscombe, G.E.M., "On the Frustration of the Majority's Will," *Analysis* 36, 4 (1976): 161-168.
- Applbaum, Arthur Isak, "Forcing a People to be Free," *Philosophy & Public Affairs* 35, 4 (2007): 359-400.
- Aristotle, *Nicomachean Ethics*, edited and translated by Roger Crisp, Cambridge: Cambridge University Press, 2000.
- Arneson, Richard, "Democracy is not Intrinsically Just," in *Justice and Democracy: Essays for Brian Barry* edited by Keith Dowding, Robert E. Goodin and Carol Pateman, Cambridge: Cambridge University Press, 2004.
- Arnold, Samuel, "Right-Wing Rawlsianism: A Critique," *Journal of Political Philosophy* (forthcoming): available online at <http://onlinelibrary.wiley.com/doi/10.1111/jopp.12001/abstract>, accessed 15 April 2013.

- Aronfreed, Justin, *Conduct and Conscience: The Socialization of Internalized Control over Behavior*, New York: Academic Press, 1968.
- Asch, Solomon, "Opinions and Social Pressure," *Scientific American* 193, 5 (1955): 31-35.
 —, "Forming Impressions of Personality," *Journal of Abnormal and Social Psychology* 41, 3 (1946): 258-290.
 —, "Studies of Independence and Conformity: A Minority of One Against a Unanimous Majority," *Psychological Monographs* 70, 9 (1956): 1-70.
- Ashford, Elizabeth, "The Demandingness of Scanlon's Contractualism," *Ethics* 113, 2 (2003): 273-302.
- Ashworth, Andrew, *Sentencing and Criminal Justice Fifth Edition*, Cambridge: Cambridge University Press, 2010.
 —, "Ignorance of the Criminal Law, and Duties to Avoid It," *The Modern Law Review* 74, 1 (2011): 1-26.
- Ashworth, Andrew and Martin Wasik (eds.), *Fundamentals of Sentencing Theory: Essays in Honour of Andrew von Hirsch*. Oxford: Clarendon Press, 1998.
- Ashworth, Andrew, Lucia Zedner and Patrick Tomlin (eds.), *Prevention and the Limits of the Criminal Law*, Oxford: Oxford University Press, 2013.
- Audi, Robert, "The Separation of Church and State and the Obligations of Citizenship," *Philosophy & Public Affairs* 18, 3 (1989): 259-96.
- Austin, John, *The Province of Jurisprudence Determined and the Uses of the Study of Jurisprudence*, Indianapolis: Hackett Publishing Company, 1998.
- Ball, Caroline, Kevin MacCormac, and Nigel Stone, *Young Offenders: Law, Policy, and Practice*, London: Sweet & Maxwell, 1995.
- Barret, Richard and Laila Bokhari, "Deradicalization and Rehabilitation Programmes Targeting Religious Terrorists and Extremists in the Muslims World: An Overview," in *Leaving Terrorism Behind: Individual and Collective Disengagement*, edited by Tore Bjørgo and John Horgan, New York: Routledge, 2009.
- Barry, Brian. "John Rawls and the Search for Stability," *Ethics* 105, 4 (1995): 874-915.
 —, *Justice as Impartiality*, New York: Oxford University Press, 1995.
 —, *Culture and Equality: An Egalitarian Critique of Multiculturalism*, Cambridge: Harvard University Press, 2001.
- Beehler, Rodger, "Containing Violence," *Ethics* 92, 4 (1982): 647-60.
- Beerbohm, Eric, *In Our Name: The Ethics of Democracy*, Princeton: Princeton University Press, 2012.
- Beitz, Charles, *Political Theory and International Relation*, Princeton: Princeton University Press, 1979.

- Benditt, Theodore M., "Liability for Failing to Rescue," *Law and Philosophy*, 1, 3 (1982): 391-418.
- Benn, Stanley, "The Problematic Rationality of Political Participation," in *Political Participation: A Discussion of Political Rationality*, Canberra: Australian National University Press, 1978.
- Benn, Stanley et al, *Political Participation: A Discussion of Political Rationality*, Canberra: Australian National University Press, 1978.
- Bentham, Jeremy, *An Introduction to the Principles of Morals and Legislation* London: University of London Press, 1970 [1789].
- , "Of the Principle of Utility," in *Bentham's Political Thought*, edited by Bikhu Parekh, London: Croom Helm Limited, 1973.
- Ben-Ze'ev, Aaron, "Envy and Jealousy," *Canadian Journal of Philosophy* 20 (1990): 487–516.
- , "Envy and Inequality," *The Journal of Philosophy* 9 (1992): 551–581.
- Berger, Rony and Philip Zimbardo, "Creating a Partner: A Qualitative Study of Political Extremists and Ex-Gang Members Who Have Chosen the Anti-violence Path," Commissioned by Google Ideas, March 2012, available at <http://ccare.stanford.edu/?article=creating-a-partner-a-qualitative-study-of-political-extremists-and-ex-gang-members-who-have-chosen-the-antiviolence-path>, accessed 15 January 2013.
- Bickel, Alexander M., *The Least Dangerous Branch: The Supreme Court at the Bar of Politics, Second Edition*, New Haven: Yale University Press, 1986.
- Bingham, John, "Gay marriage could signal return to 'centuries of persecution', say 1,000 Catholic priests," *The Telegraph*, 11 January 2013, accessed 13 January 2013 at <http://www.telegraph.co.uk/news/religion/9795680/Gay-marriage-could-signal-return-to-centuries-of-persecution-say-1000-Catholic-priests.html>
- Björge, Tore, Jaap Van Donselaar, and Sara Grunenberg, "Exit from Right-Wing Extremist Groups: Lessons from Disengagement Programmes in Norway, Sweden, in Germany," in *Leaving Terrorism Behind: Individual and Collective Disengagement*, edited by Tore Björge and John Horgan, New York: Routledge, 2009.
- Björge, Tore and John Horgan (eds.), *Leaving Terrorism Behind: Individual and Collective Disengagement*, New York: Routledge, 2009.
- Black, Jr., Charles, *A New Birth of Freedom: Human Rights, Named and Unnamed*, London and New Haven: Yale University Press, 1997.
- Blake, Michael, "Distributive Justice, State Coercion, and Autonomy," *Philosophy & Public Affairs* 30 (2001): 257–96.

- Bohman, James and William Rehg (eds.), *Deliberative Democracy: Essays on Reason and Politics*, Cambridge: MIT Press, 2002.
- Bold, Jr., Fredric J., "Vows to Collide: The Burgeoning Conflict between Religious Institutions and Same-Sex Marriage Antidiscrimination Laws," *University of Pennsylvania Law Review* 158, 1 (2009): 179-230.
- Boucek, Christopher, Shazadi Beg, and John Horgan, "Opening Up the Jihadi Debate: Yemen's Committee for Dialogue," in *Leaving Terrorism Behind: Individual and Collective Disengagement*, edited by Tore Bjørgo and John Horgan, New York: Routledge, 2009.
- Brettschneider, Corey, "Balancing Procedures and Outcomes Within Democratic Theory: Core Values and Judicial Review," *Political Studies* 53 (2005): 423-442.
- , *Democratic Rights: The Substance of Self-Government*, Princeton: Princeton University Press, 2007.
- , "Judicial Review and Democratic Authority: Absolute vs. Balancing Conceptions," *Journal of Ethics and Social Philosophy*, Symposium (2011): 1-9.
- , *When the State Speaks, What Should It Say?*, Princeton: Princeton University Press, 2012.
- Brighouse, Harry and Adam Swift, "Legitimate Parental Partiality," *Philosophy & Public Affairs* 37, 1 (2009): 43-80.
- Broome, John, "Fairness," *Proceedings of the Aristotelian Society* 91, 1 (1991): 87-101.
- Brown v. Board of Education of Topeka*, 347 U.S. 483 (1954).
- Brown, Jr., Stuart M., "Does Ought Imply Can?," *Ethics* 60, 4 (1950): 275-84.
- Brubaker, Stanley C., "Can Liberals Punish?," *The American Political Science Review* 82, 3 (1988): 821-36.
- Buchanan, James, *The Limits of Liberty*, Chicago: University of Chicago Press, 1975.
- Burgess, Anthony, *A Clockwork Orange*, Harmondsworth: Penguin Books, 1972.
- Burke, Edmund, "Speech to the Electors of Bristol," in *The Works of the Right Honourable Edmund Burke, Vol. 1*, London: Henry G. Bohn, 1854.
- , *The Works of the Right Honourable Edmund Burke, Vol. 1*, London: Henry G. Bohn, 1854.
- Burke, Jason, "Senior extremists linked to al-Qaida, surrenders to Saudi authorities," *The Guardian*, 5 May 2011, available at <http://www.guardian.co.uk/world/2011/may/05/extremist-alqaida-surrenders-saudi-arabia>, accessed 15 January 2013.
- Burns, Arnold, "Proximity and Particularism," *Ethical Perspectives* 3, 3 (1996): 157-160.
- Byrd, Sharon B., "Kant's Theory of Punishment: Deterrence in Its Threat, Retribution in Its Execution," *Law and Philosophy* 8, 2 (Aug., 1989): 151-200.

- Calhoun, Craig (ed.), *Habermas and the Public Sphere*, Cambridge: MIT Press, 1992.
- Callan, Eamonn, *Creating Citizens: Political Education and Liberal Democracy*, New York: Oxford University Press, 1997.
- Camus, Albert, *The Outsider*, translated from the French by Joseph Laredo, London: Penguin Books, 2000.
- Carr, David and Jan Steutel (eds.), *Virtue Ethics and Moral Education*, New York: Routledge, 1999.
- Chalmers, David. "Does Conceivability Entail Possibility?" in *Conceivability and Possibility*, edited by T. Gendler & J. Hawthorne, Oxford: Oxford University Press, 2002.
- Christiano, Thomas, *The Constitution of Equality: Democratic Authority and Its Limits*, Oxford: Oxford University Press, 2008.
- Citizens United v. Federal Election Commission*, 558 U.S. 310 (2010).
- Clayton, Matthew, *Justice and Legitimacy in Upbringing*, New York: Oxford University Press, 2006.
- Clark, Philip, "Kantian Morals and Humean Motives," *Philosophy and Phenomenological Research* 68, 1 (2004): 109-126.
- Cohen, G.A., *Rescuing Justice and Equality*, Cambridge: Harvard University Press, 2008.
- Cohen, Jared and Brendan Ballou, *Becoming a Former: Identity, Ideology, and Counterradicalization*, Washington, DC: Council on Foreign Relations, 2013.
- Cohen, Joshua, "Procedure and Substance in Deliberative Democracy," in *Democracy and Difference*, edited by Seyla Benhabib, Princeton: Princeton University Press, 1996.
- , "Deliberation and Democratic Legitimacy," in *Deliberative Democracy: Essays on Reason and Politics*, edited by James Bohman and William Rehg, Cambridge: MIT Press, 1997.
- , "Democratic Equality," *Ethics* 99, 4 (1998): 727-51.
- , "Minimalism About Human Rights: The Most We Can Hope For?," *The Journal of Political Philosophy* 12, 2 (2004): 190-213.
- , *The Arc of the Moral Universe and Other Essays*, Cambridge: Harvard University Press, 2010.
- , *Rousseau: A Free Community of Equals*, Oxford: Oxford University Press, 2010.
- , *Philosophy, Politics, Democracy: Selected Essays*, Cambridge: Harvard University Press, 2009.
- , "For a Democratic Society," in *Cambridge Companion to Rawls*, edited Samuel Freeman, Cambridge: Cambridge University Press, 2003.
- , "Truth and Public Reason," *Philosophy & Public Affairs* 37, 1 (2009): 2-42.
- Coleman, Jules, *Risks and Wrongs*, Cambridge: Cambridge University Press, 1992.
- (ed.), *Hart's Postscript*, Oxford: Oxford University Press, 2001.

- Collingridge, D.G., “‘Ought-Implies-Can’ and Hume’s Rule,” *Philosophy* 52, 201 (1977): 348-51.
- de Condorcet, Marquis, “Essay on the Application of Mathematics to the Theory of Decision-Making,” in *Condorcet: Selected Writings*, edited by Keith Michael Baker, London: Macmillan Publishing Company, 1976.
- Coote, Anna and Jo Lenaghan, *Citizens’ Juries: Theory into Practice*, London: Institute for Public Policy Research, 1997.
- Cunneen, Chris and Carolyn Hoyle, *Debating Restorative Justice*, Oxford: Hart Publishing, 2010.
- Dagger, Richard, “Playing Fair with Punishment,” *Ethics* 103, 3 (1993): pp. 473-488.
—, “Punishment as Fair Play,” *Res Publica* 14, 4 (2008): pp. 259-275.
- Darwall, Stephen, *The Second-Person Standpoint: Morality, Respect and Accountability*, Cambridge: Harvard University Press, 2009.
- Davis, Lawrence H., “They Deserve to Suffer,” *Analysis* 32, 4 (1972): 136-40.
- Decker, Scott H., David C. Pyrooz, *Leaving the Gang: Logging Off and Moving On*, Washington, DC: Council on Foreign Relations, 2011.
- DeLue, Steven M., “Aristotle, Kant and Rawls on Moral Motivation in a Just Society,” *American Political Science Review* 74, 2 (1980): 385-93.
- Dewey, John, *Democracy and Education: An Introduction to the Philosophy of Education*, New York: The Free Press, 1916.
- District of Columbia v. Heller*, 554 U.S. 570 (2008).
- Doris, John M., *Lack of Character: Personality and Moral Behavior*, Cambridge: Cambridge University Press, 2002.
- Dreben, Burton, “On Rawls and Political Liberalism,” in *The Cambridge Companion to Rawls*, edited by Samuel Freeman, New York: Cambridge University Press, 2003.
- Dryzek, John, *Discursive Democracy: Politics, Policy, and Political Science*, Cambridge: Cambridge University Press, 1990.
—, *Deliberative Democracy and Beyond: Liberals, Critics, Contestations*, New York: Oxford University Press, 2000.
- Dowding, Keith, Robert E. Goodin and Carol Pateman (eds.), *Justice and Democracy: Essays for Brian Barry*, Cambridge: Cambridge University Press, 2004.
- Dubber, Markus Dirk, *The Sense of Justice: Empathy in Law and Punishment*, New York: New York University Press, 2006.

- Duff, R.A., *Intention, Agency and Criminal Liability: Philosophy of Action and the Criminal Law*, Oxford: Basil Blackwell, 1990.
- , *Trials and Punishments*, Cambridge: Cambridge University Press, 1991.
- , *Punishment, Communication, and Community*, Oxford: Oxford University Press, 2003.
- , “Punishment and the Duties of Offenders,” *Law and Philosophy* 32, 1 (2013): 109-127.
- Duff, R.A. and S. E. Marshall, “Benefits, Burdens and Responsibilities: Some Ethical Dimensions of Situational Crime Prevention,” in *Ethical and Social Perspectives on Situational Crime Prevention* edited by Andrew von Hirsch, David Garland and Alison Wakefield, Portland: Hart Publishing, 2000.
- Dunn, John, *The Political Thought of John Locke: An Historical Account of the Argument of the ‘Two Treatises of Government’*, Cambridge: Cambridge University Press, 1969.
- Dworkin, Ronald, “Justice and Rights,” in *Taking Rights Seriously*, Cambridge: Harvard University Press, 1978.
- , *A Bill of Rights for Britain*, London: Chatto & Windus, 1990.
- , “Foundations of Liberal Equality,” in *The Tanner Lectures on Human Values*, volume 11, edited by G. B. Peterson, Salt Lake City: University of Utah Press, 1990.
- , *Freedom’s Law: The Moral Reading of the American Constitution*. Cambridge: Harvard University Press, 1996.
- , “The Curse of American Politics,” *New York Review of Books*, 17 October 1996.
- , *Law’s Empire*, Oxford: Hart Publishing, 1998.
- , “The Original Position,” in *What is Justice: Classic and Contemporary Readings Second Edition*, edited by Robert C. Solomon & Mark C. Murphy, New York: Oxford University Press: 2000.
- , *Sovereign Virtue: The Theory and Practice of Equality*, Cambridge: Harvard University Press, 2002.
- , “Hart’s Postscript and the Point of Political Philosophy,” in *Justice in Robes*, Cambridge: Harvard University Press, 2006.
- , “Rawls and the Law,” in *Justice in Robes*, Cambridge: Harvard University Press, 2006.
- , *Justice in Robes*, Cambridge: Harvard University Press, 2006.
- , “The Decision that Threatens Democracy,” *New York Review of Books*, 13 May 2010.
- , *Justice for Hedgehogs*, Cambridge: Harvard University Press, 2011.
- Elster, Jakob, “How Outlandish Can Imaginary Cases Be?,” *Journal of Applied Philosophy* 28, 3 (2011): pp. 241-258.
- Elster, Jon, *Ulysses and the Sirens: Studies in Rationality and Irrationality*, Cambridge: Cambridge University Press, 1979.
- , *Sour Grapes: Studies in the subversion of rationality*, Cambridge: Cambridge University Press, 1983.
- , *Ulysses Unbound: Studies in Rationality, Precommitment, and Constraints*, Cambridge: Cambridge University Press, 2000.
- Elster, Jon and Rune Slagstad (eds.), *Constitutionalism and Democracy*, Cambridge: Cambridge University Press, 1988.

- Ely, John Hart, *Democracy and Distrust: A Theory of Judicial Review*, Cambridge: Harvard University Press, 1980.
- Employment Division v. Smith*, 494 U.S. 872 (1990).
- Estlund, David M., "The Democracy/Contractualism Analogy," *Philosophy & Public Affairs* 31, 3 (2003): 387-412.
- , *Democratic Authority: A Philosophical Framework*, Princeton: Princeton University Press, 2008.
- , "Human Nature and the Limits (if any) of Political Philosophy," *Philosophy & Public Affairs*, 39, 3 (2011): 207-237.
- , "Reply to Copp, Gaus, Richardson, and Edmundson," *Ethics* 121, 2 (2011): 354-89.
- Fabre, Cecile. "Mandatory Rescue Killings," *The Journal of Political Philosophy* 15, 4 (2007): 363-384.
- Fallon, Richard, "The Core of an Uneasy Case For Judicial Review," *Harvard Law Review* 121, 7 (2008): 1693-1736.
- Farrell, Daniel, "The Justification of General Deterrence," *The Philosophical Review* 94, 3 (1985): 367-94.
- , "Punishment Without the State," *Noûs* 22, 3 (1988): 437-53.
- , "On Threats and Punishments," *Social Theory and Practice* 15, 2 (1989): 125-54.
- , "The Justification of Deterrent Violence," *Ethics* 100, 2 (1990): 301-17.
- Farrelly, Colin, "Justice in Ideal Theory: A Refutation," *Political Studies* 55 (2007): 844-64.
- Feinberg, Joel, "The Expressive Function of Punishment," in *Doing and Deserving*, Princeton: Princeton University Press, 1970.
- , *Doing and Deserving*, Princeton: Princeton University Press, 1970.
- Fine, Sarah, "Freedom of Association is Not the Answer," *Ethics* 120, 2 (2010): 338-356.
- Fink, Naureen Chowdhury, and Hamed El-Said, *Transforming Terrorists: Examining International Efforts to Address Violent Extremism*, New York: International Peace Institute, 2011.
- Finn, J.E., *Constitutions in Crisis*, New York: Oxford University Press, 1991.
- Finnis, John, *Natural Law and Natural Rights*, New York: Oxford University Press, 1980.
- Fischer, John Marin, "'Ought Implies Can', Causal Determinism and Moral Responsibility," *Analysis* 63, 3 (2003): 244-50.
- Fishkin, James S., *Democracy and Deliberation: New Directions for Democratic Reform*, New Haven: Yale University Press, 1991.
- , *The Voice of the People: Public Opinion and Democracy*, revised edition, New Haven: Yale University Press, 1997.

- Foot, Philippa, "The Problem of Abortion and the Doctrine of the Double Effect," in *Virtues and Vices and Other Essays in Moral Philosophy*, Los Angeles: UCLA Press, 1978.
- Forst, Rainer, *Contexts of Justice: Political Philosophy beyond Liberalism and Communitarianism*, translated by John M.M. Farrell, Berkley: University of California Press, 2002.
- , *The Right to Justification: Elements of a Constructivist Theory of Justice*, translated by Jeffrey Flynn, New York, Columbia University Press, 2012.
- Foucault, Michel, *Discipline and Punish*, London: Penguin Books, 1991.
- Fraser, Nancy, "Rethinking the Public Sphere," in *Habermas and the Public Sphere*, edited by Craig Calhoun, Cambridge: MIT Press, 1992.
- Freeman, Samuel. "Constitutional Democracy and the Legitimacy of Judicial Review," *Law and Philosophy* 9, 4 (1990-1991) 327-370.
- , "Illiberal Libertarians: Why Libertarianism is Not a Liberal View," *Philosophy & Public Affairs* 30, 2 (2001): 105-51.
- (ed.), *The Cambridge Companion to Rawls*, New York: Cambridge University Press, 2003.
- , "Congruence and the Good of Justice," in *Justice and the Social Contract: Essays on Rawlsian Political Philosophy*, New York: Oxford University Press, 2007.
- , *Justice and the Social Contract: Essays on Rawlsian Political Philosophy*, New York: Oxford University Press, 2007.
- French, Damien and Michael Laver, "Participation Bias, Durable Opinion Shifts and Sabotage through Withdrawal in Citizens' Juries," *Political Studies* 57, 2 (2009): 422-50.
- Fricker, Miranda, *Epistemic Injustice: Power and the Ethics of Knowing*, New York: Oxford University Press, 2007.
- Friedman, Marilyn, "John Rawls and the Political Coercion of Unreasonable People," in *Autonomy, Gender, Politics*, New York: Oxford University Press, 2003.
- , *Autonomy, Gender, Politics*, New York: Oxford University Press, 2003.
- Fuller, Lon, *The Morality of Law*, New Haven, Yale University Press, 1964.
- Galston, William, "On the Alleged Right to Do Wrong: A Response to Waldron," *Ethics* 93, 2 (1983): 320-24.
- , *Liberal Purposes*, Cambridge: Cambridge University Press, 1991.
- Galvin, Richard, "Does Kant's Psychology of Morality Need Basic Revision?," *Mind* 100, 2 (1991): 221-36.
- Gardner, Dan, *Risk: The Science and Politics of Fear*, London: Virgin Books, 2008.
- Gardner, John (ed.), *Punishment and Responsibility*, Oxford: Oxford University Press, 2008.
- , *Offences and Defences: Selected Essays in the Philosophy of Criminal Law*, New York: Oxford University Press, 2007.

- , “Can There Be a Written Constitution?” in *Oxford Studies in the Philosophy of Law*, Volume 1, edited by Leslie Green and Brian Leiter, New York: Oxford University Press, 2011.
- Garfinkel, Renee, “From Religious Extremist to Peacemaker,” *Common Ground News Service: Constructive Articles that Foster Dialogue*, 19 June 2007, available at <http://www.commongroundnews.org/article.php?sid=1&id=21200>, accessed 5 January 2013.
- Gaus, Gerald F., *Justificatory Liberalism: An Essay on Epistemology and Political Theory*, New York: Oxford University Press, 1996.
- , *The Order of Public Reason: A Theory of Freedom and Morality in a Diverse and Bounded World*, Cambridge: Cambridge University Press, 2011.
- , “Review: *Why Political Liberalism? On John Rawls’s Political Turn*,” *Ethics* 122, 1 (2011): 220-224.
- , *The Order of Public Reason*, Cambridge: Cambridge University Press, 2011.
- , “On the Appropriate Mode of Justifying a Public Moral Constitution,” *The Harvard Review of Philosophy* (forthcoming).
- , “The Turn to a Political Liberalism,” in *A Companion to Rawls*, edited by David Reidy and Jon Mandle, New York: Wiley-Blackwell, forthcoming.
- Gauthier, David, *Morals by Agreement*, Oxford: Oxford University Press, 1986.
- , “Moral Artifice,” *Canadian Journal of Philosophy* 18, 2 (1988): 385-418.
- Gendler, T. & J. Hawthorne, (eds.), *Conceivability and Possibility*, Oxford: Oxford University Press, 2002.
- Geuss, Raymond, *Philosophy and Real Politics*, Princeton: Princeton University Press, 2008.
- Golash, Deirdre, *The Case Against Punishment: Retribution, Crime Prevention, and the Law*, New York: New York University Press, 2005.
- Goodin, Robert E., *Utilitarianism as a Public Philosophy*, Cambridge: Cambridge University Press, 1995.
- , *Reflective Democracy*, New York: Oxford University Press, 2003.
- , “Disgorging the Fruits of Historical Wrongdoing,” *American Political Science Review*, forthcoming.
- Goodin, Robert, E. and John S. Dryzek, “Deliberative Impacts: Macro-Political Uptake of Mini-Publics,” *Politics & Society* 34, 2 (2006): 219-244.
- Goodin, Robert E. and David Estlund, “The Persuasiveness of Democratic Majorities,” *Philosophy, Politics & Economics* 3, 2 (2004): 131-42.
- Goodin, Robert E. and Christian List, “Epistemic Democracy: Generalizing the Condorcet Jury Theorem,” *Journal of Political Philosophy* 9, 3 (2001): 277-306.
- Gough, Stephen and Andrew Stables (eds.), *Sustainability and Security within Liberal Societies: Learning to Live with the Future*, New York: Routledge, 2008.

Graef, Roger, *Why Restorative Justice: Repairing the harm caused by crime*, edited and additional material by Marian Liebmann, London: Calouste Gulbenkian Foundation, 2001.

Green, Leslie, "Law, Co-Ordination, and the Common Good," *Oxford Journal of Legal Studies* 3, 3 (1983): 299-324.

—, *The Authority of the State*, New York: Oxford University Press, 1988.

—, "The Germ of Justice," *Oxford Legal Studies Research Papers No. 60/2010*, available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1703008, accessed 20 March 2013.

Gutmann, Amy, *Democratic Education*, Princeton: Princeton University Press, 1987.

Gutmann, Amy and Dennis Thompson, *Democracy and disagreement: Why moral conflict cannot be avoided in politics, and what should be done about it*, Cambridge: Harvard University Press, 1996.

—, *Why Deliberative Democracy?*, Princeton: Princeton University Press, 2004.

Habermas, Jürgen, *Moral Consciousness and Communicative Action*, translated by Christian Lenhardt and Shierry Weber Nicholsen, Cambridge: The MIT Press, 1990.

—, *Between Facts and Norms: Contributions to a Discourse Theory of Law and Democracy*, translated by William Rehg, Cambridge: The MIT Press, 1998.

—, "Reconciliation Through the Public Use of Reason," in *The Inclusion of the Other*, edited by Cíaran Cronin and Pablo De Greff, Cambridge: The MIT Press, 1998.

—, *The Inclusion of the Other: Studies in Political Theory*, edited by Cíaran Cronin and Pablo De Greff, Cambridge: The MIT Press, 1998.

—, "Religion in the Public Sphere," *European Journal of Philosophy* 14, 1 (2006): 1-25.

Haidt, Jonathan, *The Righteous Mind, Why Good People Are Divided by Politics and Religion*, London: Allen Lane, 2012.

Hampton, Jean, "The Moral Education Theory of Punishment," *Philosophy and Public Affairs* 13, 3 (1984): pp. 208-238.

Harman, Gilbert, "Moral Philosophy Meets Social Psychology: Virtue Ethics and the Fundamental Attribution Error," in *Explaining Value, and Other Essays in Moral Philosophy*, New York: Oxford University Press, 2000.

—, *Explaining Value, and Other Essays in Moral Philosophy*, New York: Oxford University Press, 2000.

—, "Guilt-Free Morality," in *Oxford Studies in Metaethics*, Volume 4, edited by Russ Schafer-Landau, Oxford: Oxford University Press, 2009.

Harsanyi, John, "Can the Maximin Principle Serve as the Basis for Morality? A Critique of John Rawls's Theory," *American Political Science Review* 69, 2 (1975): 594-606.

Hart, H.L.A., "Are There Any Natural Rights?," *Philosophical Review*, 64, 2 (1955): 175-91

—, "Prolegomenon to the Principles of Punishment," in *Punishment and Responsibility*, edited by John Gardner, Oxford: Oxford University Press, 2008 [1968].

—, *The Concept of Law Second Edition*, New York: Oxford University Press, 1994.

- Hegel, G. W. F., *The Philosophy of Right*, translated by T. Knox, Oxford: Oxford University Press, 1942 [1821].
- Heintz, Lawrence, "Excuses and 'Ought' Implies 'Can'," *Canadian Journal of Philosophy* 5, 3 (1975): 449-62.
- Henderson, G. P., "'Ought' Implies 'Can'," *Philosophy* 41, 156 (1966): 101-12.
- Herman, Barbara, "Mutual Aid and Respect for Persons," *Ethics* 94, 4 (1984): 577-602.
- Higgins, E.T., W.S. Rholes, and C.R. Jones, "Category Accessibility and Impression Formation," *Journal of Experimental Social Psychology* 13 (1977): 141-154.
- Hill, Jr., Thomas E., "Kant on Wrongdoing, Desert, and Punishment," *Law and Philosophy* 18, 4 (1999): pp. 407-441.
- , *Pluralism, Respect, and Justice*, Oxford: Oxford University Press, 2000.
- Hobbes, Thomas. *Leviathan*, edited by C.B. Macpherson, London: Penguin Books, 1968 [1651].
- Holmes, Stephen, "Gag Rules, or the Politics of Omission," in *Constitutionalism and Democracy*, edited by Jon Elster and Rune Slagstad, Cambridge: Cambridge University Press, 1988.
- , *Passions and Constraint*, Chicago: University of Chicago Press, 1995.
- Honig, Bonnie, *Political Theory and the Displacement of Politics*, Ithaca: Cornell University Press, 1993.
- , "Rawls on Politics and Punishment," *Political Research Quarterly* 46, 1 (1993): 99-125.
- Horkheimer, Max, "Materialismus und Moral," *Gesammelte Schriften, Band 3, Schriften 1931-36*, Frankfurt: Fischer Verlag, 2009.
- Hostein, Harvey A., Elizabeth LaKind, Gladys Frankel, and Stella Manne, "Effects of Knowledge About Remote Social Events on Prosocial Behaviour, Social Conception, and Mood," *Journal of Personality and Social Psychology* 32, 6 (1975): 1038-1046.
- Howard, Jeffrey, "Punishment, Socially Deprived Offenders, and Democratic Community," *Criminal Law and Philosophy* 7, 1 (2013): 121-136.
- , "Democracy as the Search for Justice: An Epistemic Justification," *Political Studies* (forthcoming).
- Howard-Snyder, Frances, "'Cannot' Implies 'Not Ought'," *Philosophical Studies* 130, 2 (2006): 233-246.
- Human Rights Act, 1998, c. 42, § 4(2), (6).
- Hume, David, *A Treatise of Human Nature*, second edition, edited by L. A. Selby-Bigge and P. H. Nidditch, Oxford: Oxford University Press, 1978.

- Ibsen, Malte Frøslee, "Justice as Non-Domination," (unpublished manuscript).
- Jacquette, Dale, "Moral Dilemmas, Disjunctive Obligations, and Kant's Principle That 'Ought' Implies 'Can'," *Synthese* 88, 1 (1991): 43-55.
- James, Aaron, "Constructing Justice for Existing Practice: Rawls and the Status Quo", *Philosophy and Public Affairs*, 33, 3 (Jul., 2005): 281-316.
- , "Constructivism about Practical Reasons", *Philosophy and Phenomenological Research*, 74, 2 (2007): 302-325.
- , "Mutual Assurance Problems in Global Context," in *Hobbes Today: Insights for the 21st Century*, edited by Sharon Lloyd, New York: Cambridge University Press, forthcoming.
- , "Constructivism, Moral," in *The International Encyclopedia of Ethics*, edited by Hugh LaFollette (Oxford: Blackwell, forthcoming).
- Jamieson, Dale, "Duties to the Distant: Aid, Assistance, and Intervention in the Developing World," *The Journal of Ethics* 9, 1/2 (2005): 151-170.
- Jefferson Center, "Jefferson Center: Originator of the Citizens Jury Process," <http://www.jefferson-center.org>, accessed 11 March 2013.
- Jefferson, Thomas, "The Declaration of Independence," in *The Declaration of Independence and the Constitution of the United States*, New York: Bantam, 1998.
- Joss, Simon and John Durand (eds.), *Public Participation in Science: The Role of Consensus Conferences in Europe*, London: Science Museum, 1995.
- Kagan, Shelly, *The Limits of Morality*, Oxford: Clarendon Press, 1989.
- Kalt, Stefan, "David Hume on the Motivating Effect of Moral Perception: Internalist or Externalist?," *History of Philosophy Quarterly* 22, 2 (2005): 143-160.
- Kamm, F.M., "Does Distance Matter Morally to the Duty to Rescue?," *Law and Philosophy* 19, 6 (2000): 655-681.
- Kane, Robert, "Deontic Acts, Frankfurt-Style Examples, and 'Ought' Implies 'Can'," *The Journal of Ethics* 4, 4 (2000): 357-60.
- Kant, Immanuel. *Practical Philosophy*, translated and edited by Mary Gregor, Cambridge: Cambridge University Press, 1996.
- , "On the Common Saying: That may be correct in theory, but it is of no use in practice," in *Practical Philosophy*, translated and edited by Mary Gregor, Cambridge: Cambridge University Press, 1996.
- , "The Metaphysics of Morals," in *Practical Philosophy*, translated and edited by Mary Gregor, Cambridge: Cambridge University Press, 1996.
- , "Groundwork for the Metaphysics of Morals," in *Practical Philosophy*, translated and edited by Mary Gregor, Cambridge: Cambridge University Press, 1996.

- , “Critique of Pure Practical Reason” in *Practical Philosophy*, translated and edited by Mary Gregor, Cambridge: Cambridge University Press, 1996.
- , *Religion within the Boundaries of Mere Reason And Other Writings*, edited by Allen Wood and George di Giovanni, Cambridge: Cambridge University Press, 1998.
- Kekes, John, “‘Ought Implies Can’ and Two Kinds of Morality,” *The Philosophical Quarterly* 34, 137 (1984): 459-67.
- Kelly, Erin and Lionel McPherson, “On Tolerating the Unreasonable,” *The Journal of Political Philosophy* 9, 1 (2001): 38-55.
- Kelsen, Hans, *General Theory of Law and State*, translated by Anders Wedberg, Cambridge: Harvard University Press, 1945; reprinted in Clark: The Lawbook Exchange Ltd., 2007.
- Kiesler, C., P. Mathog, P. Pool, and R. Howenstein, “Commitment and the Boomerang Effect: A Field Study,” in *The Psychology of Commitment*, edited by Charles Kiesler, New York: Academic Press, 1971.
- King, Martin Luther, “Paul’s Letter to American Christians,” Delivered at Dexter Avenue Baptist Church, Montgomery, Alabama, on 4 November 1956. *MLKP*, accessed 14 January 2013. http://mlk-kpp01.stanford.edu/index.php/encyclopedia/documentsentry/doc_pauls_letter_to_american_christians.
- Klosko, George, “Rawls’s Argument from Political Stability,” *Columbia Law Review* 94, 6 (1994): 1882-1897.
- , *Democratic Procedures and Liberal Consensus*, Oxford: Oxford University Press, 2004.
- Knight, Carl and Zofia Stemplowska (eds.), *Responsibility and Distributive Justice*, Oxford: Oxford University Press, 2011.
- Kolodny, Niko, “Love as a Valuing Relationship,” *The Philosophical Review* 112, 2 (2003): 135-89.
- Korsgaard, Christine M., “The Reasons We Can Share,” *Social Philosophy & Policy* 10, 1 (Dec., 1993): 24-51.
- , “Creating the Kingdom of Ends: Reciprocity and Responsibility,” in *Creating the Kingdom of Ends*, Cambridge: Cambridge University Press, 1996.
- , *Creating the Kingdom of Ends*, Cambridge: Cambridge University Press, 1996.
- , *The Sources of Normativity*, edited by Onora O’Neill, Cambridge: Cambridge University Press, 1996.
- , Korsgaard, Christine, “From Duty and for the Sake of the Noble: Kant and Aristotle on Morally Good Action,” in *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*, New York: Oxford University Press, 2008.
- , “Realism and Constructivism in Twentieth-Century Moral Philosophy” in *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*, New York: Oxford University Press, 2008.

- , *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*, New York: Oxford University Press, 2008.
- Kramer, Matthew H., *The Ethics of Capital Punishment: A Philosophical Investigation of Evil and its Consequences*, New York: Oxford University Press, 2011.
- Krasnoff, Larry, “Consensus, Stability, and Normativity in Rawls’s Political Liberalism,” *Journal of Philosophy* 95, 6 (1998): 269-292.
- Krause, Sharon, “Hume and the (False) Luster of Justice,” *Political Theory* 32, 5 (2004): 628-55.
- Kupperman, Joel, “Virtue in Virtue Ethics,” *The Journal of Ethics* 3, 2/3 (2009): 243-55.
- Kymlicka, Will, *Contemporary Political Philosophy: An Introduction*, Oxford: Clarendon Press, 1990.
- Landemore, Hélène, *Politics, Collective Intelligence, and the Rule of the Many*, Princeton: Princeton University Press, 2013.
- Larmore, Charles, *Patterns of Moral Complexity*, Cambridge: Cambridge University Press, 1987.
- , “Political Liberalism,” *Political Theory* 18, 3 (Aug., 1990): 339-360.
- , *The Morals of Modernity*, Cambridge: Cambridge University Press, 1996.
- , “The Moral Basis of Political Liberalism,” *The Journal of Philosophy* 96, 12 (1999): 599-625.
- , “What is Political Philosophy?,” *The Journal of Moral Philosophy*, 10 (forthcoming 2013).
- Laslett, Peter and W.G. Runciman (eds.), *Philosophy, Politics, and Society*, second series, Oxford: Blackwell, 1969.
- Lawford-Smith, Holly, “Understanding Political Feasibility,” *The Journal of Political Philosophy*, forthcoming.
- Laycock, David (ed.), *Representation and Democratic Theory*, Vancouver: UBC Press, 2004.
- Layton, Richard, “Discussion Group Report: Socrates Changes the Lives of Present-Day Prison Inmates,”
http://www.humanistsofutah.org/2000/SocratesChangesLivesPrisonInmates_DiscGrp_08-00.html, accessed on 20 November 2012.
- LeDoux, Joseph, “Emotional Memory Systems in the Brain,” *Behavioral Brain Research* 58, 1-2 (1993): 69-79.
- , *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*, New York: Simon and Schuster, 1996.
- Lenaghan, Jo, Bill New, and Elizabeth Mitchell, “Setting Priorities: Is There a Role for Citizens’ Juries?,” *British Medical Journal* 312, 7046 (1996): 1591-4.

- Leopold, David and Marc Stears, *Political Theory: Methods and Approaches*, Oxford: Oxford University Press, 2008.
- Levinson, Meira, *The Demands of Liberal Education*, New York: Oxford University Press, 1999.
- . *No Citizen Left Behind*, Cambridge: Harvard University Press, 2012.
- Levinson, Sanford, *Constitutional Faith*, Princeton: Princeton University Press, 1988.
- Leyden, T. J. and M. Bridget Cook, *Skinhead Confessions: From Hate to Hope*, Springville: Cedar Fort, 2008.
- Lippke, Richard L., "Mixed Theories of Punishment and Mixed Offenders: Some Unresolved Tensions," *Southern Journal of Philosophy* 44, 2 (2006): 273-95.
- Lister, Andrew, "Reciprocity, Relationships, and Distributive Justice," *Social Theory and Practice* 39, 1 (2013): 70-94.
- Lloyd, Sharon (ed.), *Hobbes Today: Insights for the 21st Century*, New York: Cambridge University Press, forthcoming.
- Lochner v. New York*, 198 U.S. 45 (1905).
- Locke, John, *Second Treatise of Government*, edited by C.B. Macpherson, Indianapolis: Hackett Publishing Company, 1980.
- Macedo, Stephen, *Liberal Virtues: Citizenship, Virtue, and Community in Liberal Constitutionalism*, Oxford: Clarendon Press, 1990.
- , *Diversity and Distrust: Civic Education in a Multicultural Democracy*, Cambridge: Harvard University Press, 2000.
- Macedo, Stephen and Yael Tamir (eds.), *NOMOS XLIII: Moral and Political Education*, New York: New York University Press, 2002.
- Mack, Eric, "Bad Samaritanism and the Causation of Harm," *Philosophy & Public Affairs* 9, 3 (1980): 230-259.
- Madison, James. "Federalist #51," *Independent Journal*, 6 February 1788, reprinted at <http://www.constitution.org/fed/federa51.htm>, accessed on 20 February 2013.
- Malm, H. M., "Liberalism, Bad Samaritan Law, and Legal Paternalism," *Ethics* 106, 1 (October, 1995): 4-31.
- March, Andrew, *Islam and Liberal Citizenship: The Search for Overlapping Consensus*, Oxford: Oxford University Press, 2009.
- Mason, Elinor, "Consequentialism and the 'Ought Implies Can' Principle," *American Philosophical Quarterly* 40, 4 (2003): 319-331.

- Matravers, Matt, *Justice and Punishment: The Rationale of Coercion*, New York: Oxford University Press, 2000.
- , “Mad, Bad or Faulty? Desert in Distributive and Retributive Justice,” in *Responsibility and Distributive Justice*, edited by Carl Knight and Zofia Stemplowska, New York: Oxford University Press, 2011.
- Mayhew, David, *Congress: The Electoral Connection*, New Haven: Yale University Press, 1974.
- McDermott, Daniel, “The Permissibility of Punishment,” *Law and Philosophy* 20, 4 (2001): 403-432.
- , “Debts to Society,” *The Journal of Political Philosophy* 10, 4 (2002): 439-64.
- , “Analytical Political Philosophy,” in *Political Theory: Methods and Approaches*, edited by David Leopold and Marc Stears, Oxford: Oxford University Press, 2008.
- McDowell, John, “Virtue and Reason,” *Monist* 62 (1979): 331–50.
- McIvor, Gill, *Sentenced to Service: The Operation and Impact of Community Service by Offenders*, Aldershot, Avebury, 1992.
- McMahan, Jeff, *Killing in War*, Oxford: Clarendon Press, 2009.
- Melissaris, Emmanuel, “Toward a Political Theory of Criminal Law: A Critical Rawlsian Account,” *New Criminal Law Review* 15, 1 (2012): 122-55.
- Merritt, Maria, “Virtue Ethics and Situationist Personality Psychology,” *Ethical Theory and Moral Practice* 3, 4 (2000): 365-83.
- Metz, Thaddeus, “How to Reconcile Liberal Politics with Retributive Punishment,” *Oxford Journal of Legal Studies* 27, 4 (2007): 683-705.
- Milgram, Stanley, *Obedience to Authority: An Experimental View*, New York: Harper and Row, 1974.
- Mill, J.S., *On Liberty*, edited by Elizabeth Rapaport, Indianapolis: Hackett Publishing Company, 1978 [1859].
- , “A Few Words on Non-Intervention,” in *Collected Works of John Stuart Mill: Essays on Equality, Law, and Education*, vol. XXI, edited by John M. Robson, Toronto: University of Toronto Press, 1984.
- , *Collected Works of John Stuart Mill: Essays on Equality, Law, and Education*, vol. XXI, edited by John M. Robson, Toronto: University of Toronto Press, 1984.
- , “Utilitarianism,” in *Utilitarianism and Other Essays*, edited by Alan Ryan, London: Penguin Books, 1987.
- Miller, Christian, “Social Psychology and Virtue Ethics,” *The Journal of Ethics* 7, 4 (2003): 365-92.
- Miller, David. *On Nationality*, Oxford: Clarendon Press, 1995.

- . “Political Philosophy for Earthlings,” in *Political Theory: Methods and Approaches*, edited by David Leopold and Marc Stears, Oxford: Oxford University Press, 2008.
- . “Taking Up the Slack? Responsibility and Justice in Situations of Partial Compliance?” in *Responsibility and Distributive Justice*, edited by Carl Knight and Zofia Stemplowska, Oxford: Oxford University Press, 2011.
- . “A Tale of Two Cities; Or, Political Philosophy as Lamentation,” in *Justice for Earthlings*. Cambridge: Cambridge University Press, 2013.
- . *Justice for Earthlings: Essays in Political Philosophy*, Cambridge: Cambridge University Press, 2013.

- Miller, Richard, “Beneficence, Duty, and Distance,” *Philosophy & Public Affairs* 32, 4 (2004): 357-383.

- Mills, Charles W., “‘Ideal Theory’ as Ideology,” *Hypatia* 20, 3 (2005): 165-84.

- Montefiore, Alan, “‘Ought’ and ‘Can,’” *The Philosophical Quarterly* 8, 30 (1958): 24-40.

- Moore, Michael S., *Placing Blame: A Theory of the Criminal Law*, Oxford: Oxford University Press, 1997.

- Morris, Herbert, “Persons and Punishment,” *The Monist* 52, 4 (1968): 475-501, p. 478.
- , “A Paternalistic Theory of Punishment,” *American Philosophical Quarterly* 18, 4 (1981): 263-271.

- Moscovici, Serge, “Toward a Theory of Conversion Behavior,” in *Advances in Experimental Social Psychology*, Volume 13, edited by Leonard Berkowitz, New York: Academic Press, 1980.

- Mulhall, Stephen and Adam Swift, *Liberals and Communitarians*, second edition, Oxford: Blackwell Press, 1996.

- Murphy, Liam, “Institutions and the Demands of Justice,” *Philosophy & Public Affairs* 27, 4 (1999): 251-91.
- , *Moral Demands in Nonideal Theory*, New York: Oxford University Press, 2000.

- Murphy, Mark C., “Acceptance of Authority and the Duty to Comply with Just Institutions: A Comment on Waldron,” *Philosophy & Public Affairs* 23, 3 (1994): 271-277.

- Nagel, Thomas. *The Possibility of Altruism*, Princeton: Princeton University Press, 1970.
- , “Moral Luck,” in *Mortal Questions*, New York: Cambridge University Press, 1979.
- , *Mortal Questions*, New York: Cambridge University Press, 1979.
- , *The View From Nowhere*, New York: Oxford University Press, 1986.
- , “Moral Conflict and Political Legitimacy,” *Philosophy & Public Affairs* 16 (1987): 215-40.
- , *Equality and Partiality*, New York: Oxford University Press, 1991.
- , “The Problem of Global Justice,” *Philosophy & Public Affairs*, 33, 2 (2005): 113-147.

- National Federation of Independent Business v. Sebelius*, 567 U.S. (2012).

- Nauckhoff, Josefine, "Incentives and Interests in Kant's Moral Psychology," *History of Philosophy Quarterly* 20, 1 (2003): 41-60.
- Nawaz, Maajid, *Radical: My Journey from Islamic Extremism to a Democratic Awakening*, London: W. H. Allen, 2012.
- Nawaz, Maajid and Timothy Zaal, "The Road to Damascus," *Global: The International Briefing* 9 (2012), available at <http://www.global-briefing.org/2012/01/the-road-to-damascus>, accessed 10 January 2013.
- Nelson, William, *On Justifying Democracy*, London: Routledge and Kegan Paul, 1980.
- Nietzsche, Friedrich. *On the Genealogy of Morality*, translated by Carol Diethe, Cambridge: Cambridge University Press, 1994.
- Nozick, Robert, *Anarchy, State, and Utopia*, New York: Basic Books, 1974.
- Nussbaum, Martha C., *The Fragility of Goodness: Luck and Ethics in Greek Tragedy and Philosophy*, Cambridge: Cambridge University Press, 1986.
- , *Cultivating Humanity*, Cambridge: Harvard University Press, 1997.
- , "Virtue Ethics: A Misleading Category?," *The Journal of Ethics* 3, 3 (1999): 163-201.
- , *Women and Human Development: The Capabilities Approach*, Cambridge: Cambridge University Press, 2000.
- , *Frontiers of Justice: Disability, Nationality, Species Membership*, Cambridge: Harvard University Press, 2006.
- , *Liberty of Conscience: In Defense of America's Tradition of Religious Equality*, New York: Basic Books, 2008.
- , *Not For Profit: Why Democracy Needs the Humanities*, Princeton: Princeton University Press, 2010.
- , "Political Liberalism and Perfectionist Liberalism," *Philosophy & Public Affairs* 39, 1 (2011): 3-45.
- , *The New Religious Intolerance: Overcoming the Politics of Fear in an Anxious Age*, Cambridge: Harvard University Press, 2012.
- Okin, Susan Moller, *Justice, Gender and the Family*, New York: Basic Books: 1989.
- O'Neill, Onora, "Abstraction, Idealization and Ideology in Ethics," in *Moral Philosophy and Contemporary Problems*, edited by D.G. Evans, Cambridge: Cambridge University Press, 1988.
- , *Towards Justice and Virtue*, Cambridge: Cambridge University Press, 1996.
- Otsuka, Michael, "Quinn on Punishment and Using Persons as Means," *Law and Philosophy* 15, 2 (1996): 201-08.
- , *Libertarianism Without Inequality*, New York: Oxford University Press, 2003.
- Parekh, Bikhu (ed.), *Bentham's Political Thought*, London: Croom Helm Limited, 1973.
- Parfit, Derek, *Reasons and Persons*, Oxford: Oxford University Press, 1984.

- , “Equality and Priority,” *Ratio* 10, 3 (1997): 202-21.
- , *On What Matters*, Volume 1, New York: Oxford University Press, 2011.
- People v. Decina*, 1 A.2d 592 (N.Y. Ct. App. 1956), Section 3.
- Persson, Ingmar, “Hume—Not a ‘Humean’ about Motivation,” *History of Philosophy Quarterly*, 14, 2 (1977): 189–206.
- Pettit, Phillip, *A Theory of Freedom: From the Psychology to the Politics of Agency*, New York: Oxford University Press, 2001.
- Phillips, Michael, “Reflections on the Transition from Ideal to Non-Ideal Theory,” *Noûs* 19, 4 (1985): 551-70.
- Pinker, Steven, *The Better Angels of Our Nature*, London: Penguin Books, 2011.
- Planned Parenthood of Se. Pa. v. Casey, 505 U.S. 833 (1992)
- Plato, *The Republic*, second edition, translated by Allan Bloom, New York: Basic Books, 1968.
- , “Apology” and “Crito,” in *Five Dialogues*, translated by G. M. A. Grube, Indianapolis: Hackett Publishing Company, 2002.
- Quinn, Warren, “The Right to Threaten and the Right to Punish,” *Philosophy & Public Affairs* 14, 4 (1985): 327-373.
- Quong, Jonathan, “The Rights of Unreasonable Citizens,” *The Journal of Political Philosophy* 12, 3 (2004): 314-335.
- , “What Do Citizens Need to Share?” in *Representation and Democratic Theory*, edited by David Laycock, Vancouver: UBC Press, 2004.
- , “Disagreement, Asymmetry, and Liberal Legitimacy,” *Politics, Philosophy & Economics* 4, 3 (2005): 301-330.
- , *Liberalism Without Perfection*, New York: Oxford University Press, 2011.
- Rabasa, Angel, Stacie Pettyjohn, Jeremy Ghez, Christopher Boucek, *Deradicalizing Islamist Extremists*, Arlington: RAND Corporation, 2010.
- Radcliffe, Elizabeth S., “Hume on Motivating Sentiments, the General Point of View, and the Inculcation of ‘Morality’,” *Hume Studies* 20, 1 (1994): 37-58.
- Ramzi, Walid, “Algerian Imams Engage Militants in Dialogue,” *Magharebia*, 24 August 2012, available at http://magharebia.com/en_GB/articles/awi/features/2012/08/24/feature-02, accessed on 15 January 2013.
- Rawls, John, “Two Concepts of Rules,” *Philosophical Review* 64 (1) (1955): pp. 4-13.
- , *A Theory of Justice*, Revised Edition, Cambridge: Harvard University Press 1999 [1971].
- , Kantian Constructivism in Moral Theory,” *Journal of Philosophy* 77, 9 (1980): 515-72.
- , *Political Liberalism, expanded edition*, New York: Columbia University Press, 1993.

- , “Distributive Justice: Some Addenda,” in *Collected Papers*, edited by Samuel Freeman, Cambridge: Harvard University Press, 1999.
- , “Outline of a Decision Procedure for Ethics,” in *Collected Papers*, edited by Samuel Freeman, Cambridge: Harvard University Press, 1999.
- , “The Domain of the Political and Overlapping Consensus,” in *Collected Papers*, edited by Samuel Freeman, Cambridge: Harvard University Press, 1999.
- , “The Independence of Moral Theory,” in *Collected Papers*, edited by Samuel Freeman, Cambridge: Harvard University Press, 1999.
- , *Collected Papers*, edited by Samuel Freeman, Cambridge: Harvard University Press, 1999.
- , *Lectures on the History of Moral Philosophy*, edited by Barbara Herman, Cambridge: Harvard University Press, 2000.
- , “The Idea of Public Reason Revisited,” in *The Law of Peoples*, Cambridge: Harvard University Press, 2001.
- , *The Law of Peoples*, Cambridge: Harvard University Press, 2001.
- , *Justice as Fairness: A Restatement*, edited by Erin Kelly, Cambridge: Harvard University Press, 2001.
- , *Lectures on the History of Political Philosophy*, edited by Samuel Freeman, Cambridge: Harvard University Press, 2007.
- , *A Brief Inquiry into the Meaning of Sin & Faith with “On My Religion”*, edited by Thomas Nagel, Cambridge: Harvard University Press, 2009.

- Raz, Joseph. “Authority and Justification,” *Philosophy & Public Affairs* 14, 1 (1985), 3-29.
- , *The Morality of Freedom*, Oxford: Clarendon Press, 1986.
- , *Practical Reason and Norms*, Oxford: Oxford University Press, 1999.

- Reath, Andrew, “Kant’s Theory of Moral Sensibility,” in *Agency and Autonomy in Kant’s Moral Philosophy: Selected Essays*, Oxford: Oxford University Press, 2006.

- Rehfeld, Andrew, “Offensive Political Theory,” *Perspectives on Politics* 8, 2 (2010): 465-86.

- Reidy, David and Jon Mandle (eds.), *Blackwell Companion to Rawls*, New York: Wiley-Blackwell, forthcoming.

- Ridge, Michael, “Saving Scanlon: Contractualism and Agent-Relativity,” *Journal of Political Philosophy* 9, 4 (2001): 472-81.

- Ripstein, Arthur, *Force and Freedom: Kant’s Legal and Political Philosophy*, Cambridge: Harvard University Press, 2009.

- Roach, Kent, “Common Law Bills of Rights as Dialogue Between Courts and Legislatures,” *The University of Toronto Law Journal* 55, 3 (2005): 733-66.

- Robertson, David, *The Judge as Political Theorist: Contemporary Constitutional Review*, Princeton: Princeton University Press, 2010.

- Robinson, Anne, *Foundations for Offender Management: Theory, Law and Philosophy for Contemporary Practice*, Bristol: The Policy Press, 2011.

- Robinson, Gwen, "Late-modern Rehabilitation: The Evolution of a Penal Strategy," *Punishment and Society* 10, 4 (2008): 429-445.
- Rorty, Richard, *Contingency, Irony, and Solidarity*, New York: Cambridge University Press, 1989.
- , "The Priority of Democracy to Philosophy," in *Objectivity, Relativism, and Truth: Philosophical Papers*, Cambridge: Cambridge University Press, 1991.
- , *Objectivity, Relativism, and Truth: Philosophical Papers*, Cambridge: Cambridge University Press, 1991.
- Rosati, Connie, S., "Moral Motivation," *The Stanford Encyclopedia of Philosophy* (Fall 2008 Edition), Edward N. Zalta (ed.), available at <http://plato.stanford.edu/archives/fall2008/entries/moral-motivation>, accessed 25 December 2012.
- Rosberg, Gerald M., "Aliens and Equal Protection: Why Not the Right to Vote?," *Michigan Law Review* 75, 5/6 (1977): 1092-1136.
- Rosenblum, Nancy, *Membership & Morals: the Personal Uses of Pluralism in America*, Princeton: Princeton University Press, 1998.
- , *On the Side of the Angels: An Appreciation of Parties and Partisanship*, Princeton: Princeton University Press, 2008.
- Rousseau, Jean-Jacques, *Emile or On Education*, translated by Allan Bloom, New York: Basic Books, 1979.
- , "On the Social Contract," in *The Basic Political Writings*, translated by Donald A. Cress, Indianapolis: Hackett Publishing Company, 1987.
- , "Discourse on the Origin of Inequality," in *The Basic Political Writings*, translated by Donald A. Cress, Indianapolis: Hackett Publishing Company, 1987.
- , *The Basic Political Writings*, translated by Donald A. Cress, Indianapolis: Hackett Publishing Company, 1987.
- Ryan, Alan (ed.), *Utilitarianism and Other Essays*, London: Penguin Books, 1987.
- Sager, Lawrence G., "The Why of Constitutional Essentials," *Fordham Law Review* 72, 5 (2004): 1421-1433.
- Said, Rajeh, "Why al-Qaeda members defect," *Magharebia*, 24 November 2011, available at http://magharebia.com/en_GB/articles/awi/features/2011/11/24/feature-02, accessed on 15 January 2013.
- Sandel, Michael, *Liberalism and the Limits of Justice*, second edition, Cambridge: Cambridge University Press, 1998.
- Sangiovanni, Andrea. "The Irrelevance of Coercion, Imposition, and Framing to Distributive Justice," *Philosophy & Public Affairs* 40, 2 (2012): 79-110.
- , "Global Justice, Reciprocity, and the State," *Philosophy & Public Affairs* 35, 1 (2007): 3-39.

- Saka, Paul, "Ought Does Not Imply Can," *American Philosophical Quarterly* 37, 2 (2000): 93-105.
- Scalia, Antonin, *A Matter of Interpretation: Federal Courts and the Law*, Princeton: Princeton University Press, 1997.
- Scanlon, T.M., "Contractualism and Utilitarianism," in *Utilitarianism and Beyond*, edited by Amartya Sen and Bernard Williams, Cambridge: Cambridge University Press, 1982.
- , *What We Owe to Each Other*, Cambridge: Harvard University Press, 1998.
- , "A Theory of Freedom of Expression," in *The Difficulty of Tolerance: Essays in Political Philosophy*, Cambridge: Cambridge University Press, 2003.
- , "Freedom of Expression and Categories of Expression," in *The Difficulty of Tolerance: Essays in Political Philosophy*, Cambridge: Cambridge University Press, 2003.
- , *The Difficulty of Tolerance: Essays in Political Philosophy*, Cambridge: Cambridge University Press, 2003.
- , *Moral Dimensions: Permissibility, Meaning, Blame*, Cambridge: Harvard University Press, 2008.
- , "Giving Desert Its Due," *Philosophical Explorations* (forthcoming).
- Schauber, Nancy, "Hume on Moral Motivation: It's Almost like Being in Love," *History of Philosophy Quarterly* 16, 3 (1999): 341-366.
- Scheffler, Samuel, "Justice and Desert in Liberal Theory," in *Boundaries and Allegiances: Problems of Justice and Responsibility in Liberal Thought*, New York: Oxford University Press, 2001.
- , *Boundaries and Allegiances: Problems of Justice and Responsibility in Liberal Thought*, New York: Oxford University Press, 2001.
- , "What Is Egalitarianism?," *Philosophy & Public Affairs* 31, 1 (2003): 5-39.
- , "Projects, Relationships, and Reasons," in *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, edited by R. Jay Wallace, et al., Oxford: Clarendon Press, 2006.
- Schemmel, Christian, "Why Relational Egalitarians Should Care About Distributions," *Social Theory and Practice* 37, 3 (2011): 365-90.
- Schmidtz, David, "Islands in a Sea of Obligation: Limits of the Duty to Rescue," *Law and Philosophy* 19, 6 (2000): 683-705.
- Schnall, Ira M., "The Principle of Alternate Possibilities and 'Ought' Implies 'Can'," *Analysis* (2001): 335-40.
- Schwartzman, Micah, "Judicial Sincerity," *Virginia Law Review* 94, 4 (2008): 987-1027.
- , "The Sincerity of Public Reason," *The Journal of Political Philosophy* 19, 4 (2011): 375-98.
- , "The Ethics Reasoning from Conjecture," *The Journal of Moral Philosophy* 9 (2012): 521-44.
- Sclove, Richard E., "Town Meetings on Technology: Consensus Conferences," <http://www.loka.org/pubs/techrev.htm>, accessed 11 March 2013.

- Sen, Amartya, *The Idea of Justice*, London: Allen Lane, 2009.
- Sen, Amartya and Bernard Williams (eds.), *Utilitarianism and Beyond*, Cambridge: Cambridge University Press, 1982.
- Shafter-Landau, Russ, *Moral Realism: A Defence*, Oxford: Clarendon Press, 2003.
- Sher, George, Desert, Princeton: Princeton University Press, 1989.
- , *Who Knew? Responsibility Without Awareness*, New York: Oxford University Press, 2009.
- Sherbert v. Verner*, 374 U.S. 398 (1963).
- Shklar, Judith, *The Faces of Injustice*, New Haven: Yale University Press, 1990.
- Shue, Henry, *Basic Rights: Subsistence, Affluence and US Foreign Policy*, second edition, Princeton: Princeton University Press, 1996.
- Simmons, A. John, *The Lockean Theory of Rights*, Princeton: Princeton University Press, 1994.
- , “Ideal and Nonideal Theory,” *Philosophy & Public Affairs* 38, 1 (2010): 5-36.
- Singer, Peter, “Famine, Affluence, and Morality,” *Philosophy and Public Affairs* 1, 1 (1972): 229-243.
- Sinnott-Armstrong, Walter, “‘Ought’ Conversationally Implies ‘Can’,” *The Philosophical Review* 93, 2 (1984): 249-261.
- Slote, Michael, “Virtue Ethics and Democratic Values,” *Journal of Social Philosophy*, 24, 2 (1993): 5–37.
- , *Morals from Motives*, Oxford: Oxford University Press, 2001.
- , *Moral Sentimentalism*, New York: Oxford University Press, 2010.
- , *The Impossibility of Perfection: Aristotle, Feminism, and the Complexities of Ethics*, New York: Oxford University Press, 2011.
- Smith, Adam, *A Theory of Moral Sentiments*, Indianapolis: Liberty Classics, 1982.
- Smith, Alan. “Locke and key,” *The Guardian*, 20 January 2004, accessed 22 February 2013 at <http://www.guardian.co.uk/education/2004/jan/20/furthereducation.uk2>.
- Snowiss, Sylvia, *Judicial Review and the Law of the Constitution*, New Haven: Yale University Press, 1990.
- Solomon, Robert C. & Mark C. Murphy (eds.), *What is Justice: Classic and Contemporary Readings*, second edition, New York: Oxford University Press: 2000.
- Southwood, Nicholas, *Contractualism and the Foundations of Morality*, New York: Oxford University Press, 2010.

- de Spinoza, Benedict, *The Chief Works of Benedict de Spinoza*, revised edition, translated by R. H. M. Elwes, London: George Bell and Sons, 1891.
- Spencer, Herbert, *The Man Versus the State*, Indianapolis: Liberty Fund, 1992.
- Sreenivasan, Gopal, "What is Non-Ideal Theory?," in *Transitional Justice: NOMOS LI* (New York: New York University Press, 2012).
- Stemplowska, Zofia, "What's Ideal About Ideal Theory?" *Social Theory and Practice* 34, 3 (2008): 319-40.
- , "Responsibility and Respect: Reconciling Two Egalitarian Visions," in *Responsibility and Distributive Justice*, edited by Carl Knight and Zofia Stemplowska, New York: Oxford University Press, 2011.
- Stern, Jessica, "Mind Over Martyr: How to Deradicalize Islamist Extremists," *Foreign Affairs* 89, 1 (2010): 95-108.
- Stern, Robert, "Does Ought Imply Can? And Did Kant Think It Does?," *Utilitas* 16, 1 (2004): 42-61.
- Stichter, Matt, "Virtues, Skills, and Right Action", *Ethical Theory and Moral Practice* 14, 1 (2011): 73–86.
- Stilz, Anna, *Liberal Loyalty: Freedom, Obligation, and the State*, Princeton: Princeton University Press, 2009.
- Stohr, Karen, "Kantian Beneficence," *The Journal of Moral Philosophy* 8, 1 (2011): 45–67.
- Stratton-Lake, Philip, *Kant, Duty, and Moral Worth*, London: Routledge, 2000.
- Strawson, P.F., "Freedom and Resentment," in *Freedom and Resentment and Other Essays*, London: Methuen, 1974.
- Sunstein, Cass R., "Constitutionalism and Secession," *University of Chicago Law Review* 58 (1991): 633-70.
- , *One Case at a Time: Judicial Minimalism on the Supreme Court*, Cambridge: Harvard University Press, 2001.
- , *The Second Bill of Rights*, New York: Basic Books, 2004.
- , *Infotopia: How Many Minds Produce Knowledge*, New York: Oxford University Press, 2006.
- , "The Most Mysterious Rights," *The New Republic*, 18 November 2007.
- , *Going to Extremes: How Like Minds Unite and Divide*, New York: Oxford University Press, 2009.
- Surowiecki, James, *The Wisdom of Crowds: Why the Many are Smarter than the Few and How Collective Wisdom Shapes Business*, New York: Doubleday, 2004.
- Sutherland, Stuart, *Irrationality*, London: Pinter and Martin, 2007.

- Swaine, Lucas, "Demanding Deliberation: Political Liberalism and the Inclusion of Islam," *Journal of Islamic Law and Culture* 11 (2009): 92-110.
- Tadros, Victor. *The Ends of Harm*, New York: Oxford University Press, 2011.
- Taurek, John M., "Should the Numbers Count?" *Philosophy & Public Affairs* 6, 4 (1977): 293-316.
- Tavris, Carol and Elliot Aronson, *Mistakes Were Made (But Not by Me): Why we justify foolish beliefs, bad decisions and hurtful acts*, London: Harcourt, 2007.
- Thaler, Richard and Cass R. Sunstein, *Nudge: Improving decisions about health, wealth and happiness*, London: Penguin Books, 2008.
- Thayer, James B., "The Origins and Scope of the American Doctrine of Constitutional Law," *Harvard Law Review* 7, 3 (1893): 129-56.
- Thompson, Dennis F., *Ethics in Congress: From Individual to Institutional Corruption*, Washington, DC: Brookings Press, 1995.
- Thomson, Judith Jarvis, "Killing, Letting Die, and the Trolley Problem," *The Monist* 59, 2 (1976): 204-17.
- , *The Realm of Rights*, Cambridge: Harvard University Press, 1990.
- Tomasi, John, "Civic Education and Ethical Subservience: From *Mozart* to *Santa Fe* and Beyond," in *NOMOS XLIII: Moral and Political Education*, edited by Stephen Macedo and Yael Tamir, New York: New York University Press, 2002.
- , *Free Market Fairness*, Princeton: Princeton University Press, 2012.
- Tomlin, Patrick, "Envy, Facts, and Justice: A Critique of the Treatment of Envy in Justice as Fairness," *Res Publica* 14, 2 (2008): 101-16.
- , "Should We Be Utopophobes about Democracy in Particular?," *Political Studies Review* 10, 1 (2012): 36-47.
- Tushnet, Mark, *Taking the Constitution Away from the Courts*, Princeton: Princeton University Press, 1999.
- Valentini, Laura, *Justice in a Globalized World: A Normative Framework*, Oxford: Oxford University Press, 2011.
- de Vise, Daniel, "College student's philosophy program brings Plato and Buddha to a Md. prison," *The Washington Post*, 1 September 2011, available at http://articles.washingtonpost.com/2011-09-01/local/35273403_1_maximum-security-prison-paper-clip-inmates, accessed 5 November 2012.
- Von Hirsch, Andrew, *Past or Future Crimes: Deservedness and Dangerousness in the Sentencing of Criminals*, Manchester: Manchester University Press, 1985.
- , *Censure and Sanctions*, New York: Oxford University Press, 1993.

- Von Hirsch, Andrew, David Garland and Alison Wakefield (eds.), *Ethical and Social Perspectives on Situational Crime Prevention*, Portland: Hart Publishing, 2000).
- Voorhoeve, Alex, *Conversations on Ethics*, New York: Oxford University Press, 2009.
- Voorhoeve, Alex and T. M. Scanlon, "The Kingdom of Ends on the Cheap," in *Conversations on Ethics*, edited by Alex Voorhoeve, New York: Oxford University Press, 2009.
- Wachtel, Ted, "My Three Decades of Using Restorative Practices with Delinquent and At-Risk Youth: Theory, Practice and Research Outcomes," A paper presented at the First World Congress on Restorative Juvenile Justice, Lima, Peru, 5 November 2009, available at [http://www.unicef.org/tdad/2tedwachtel\(1\).pdf](http://www.unicef.org/tdad/2tedwachtel(1).pdf), accessed 15 November 2012.
- Waldron, Jeremy, "Theoretical Foundations of Liberalism," in *Liberal Rights: Collected Papers 1981-1991*, Cambridge: Cambridge University Press, 1993.
- , "A Right to Do Wrong," in *Liberal Rights* (Cambridge: Cambridge University Press, 1993).
- , "Special Ties and Natural Duties," *Philosophy & Public Affairs* 22, 1 (1993): 3-30.
- , *Liberal Rights: Collected Papers 1981-1991*, Cambridge: Cambridge University Press, 1993.
- , *Law and Disagreement*, New York: Oxford University Press, 1999.
- , "Normative (or Ethical) Positivism," in *Hart's Postscript*, edited by Jules Coleman, Oxford: Oxford University Press, 2001.
- , *God, Locke, and Equality: Christian Foundations of John Locke's Political Thought*, Cambridge: Cambridge University Press, 2002.
- , "The Core of the Case Against Judicial Review," *The Yale Law Journal* 115, 6 (2006): 1346-1406.
- , "Judges as Moral Reasoners," *International Journal of Constitutional Law* 7, 1 (2009): 2-24.
- , *Torture, Terror and Trade-Offs: Philosophy for the White House*, New York: Oxford University Press, 2010.
- , "The Ethics of Engaging with Religious Contributions in Public Deliberation," *New York University School of Law Public Law and Legal Theory Research Paper Series*, Working Paper No. 10-84, 2010. Available at <http://ssrn.com/abstract=1708113>. Accessed 5 May 2011.
- , *The Harm of Hate Speech*, Cambridge: Harvard University Press, 2012.
- Wallace, George, *1963 Inaugural Address of Governor George C. Wallace*, Alabama Department of Archives and History. Available at <http://digital.archives.alabama.gov/cdm/singleitem/collection/voices/d/2952/rec/5>, accessed 15 February 2013.
- Wallace, R., Jay, et al (eds.), *Reason and Value: Themes from the Moral Philosophy of Joseph Raz*, Oxford: Clarendon Press, 2006.
- Weber, Max, "Politics as a Vocation," in *From Max Weber: Essays in Sociology*, edited and translated by H. H. Gerth and C. Wright Mills, New York: Oxford University Press, 1946.

- , “Science as a Vocation,” in *From Max Weber: Essays in Sociology*, edited and translated by H. H. Gerth and C. Wright Mills, New York: Oxford University Press, 1946.
- Weber, Michael, “More on the Motive of Duty,” *The Journal of Ethics* 11, 1 (2007): 65-86.
- Weithman, Paul. *Why Political Liberalism? On John Rawls’s Political Turn*, Oxford: Oxford University Press, 2010.
- Wellman, Christopher Heath, “Immigration and Freedom of Association,” *Ethics* 119, 1 (2008): 109-141.
- , “The Rights Forfeiture Theory of Punishment,” *Ethics* 122, 2 (2012): 371-393.
- Welsh, Brandon C. and David P. Farrington (eds.), *The Oxford Handbook of Crime Prevention*, New York: Oxford University Press, 2012.
- Wenar, Leif, “Political Liberalism: An Internal Critique,” *Ethics* 106, 1 (1995): 32-62.
- White, Jonathan and Lea Ypi, “On Partisan Political Justification,” *American Political Science Review* 105, 2 (2011): 381-396.
- Williams, Andrew, “Incentives, Inequality, and Publicity,” *Philosophy & Public Affairs* 27, 3 (1998), pp. 225-247.
- Williams, Bernard, “Internal and External Reasons,” in *Moral Luck*, Cambridge: Cambridge University Press, 1981.
- , *Moral Luck*, Cambridge: Cambridge University Press, 1981.
- , *In the Beginning Was the Deed*, Princeton: Princeton University Press, 2007.
- Wolff, Jonathan, “Fairness, Respect, and the Egalitarian Ethos,” *Philosophy & Public Affairs* 27, 2 (1998): 97-122.
- Wollheim, Richard, “A Paradox in the Theory of Democracy,” in *Philosophy, Politics, and Society*, second series, edited by Peter Laslett and W.G. Runciman, Oxford: Blackwell, 1969.
- Yack, Bernard, *Nationalism and the Moral Psychology of Community*, Chicago: University of Chicago Press, 2012.
- Zimbardo, Philip, *The Lucifer Effect: How Good People Turn Evil*, Reading: Ebury Publishing, 2007.